Title

Unveiling the Human Nasopharyngeal Microbiome Compendium: Systematic Characterization of Community Architecture and Function Through a Comprehensive Meta-Analysis

Running Title

Meta-Analysis of Human Nasopharyngeal Microbiome Communities

Authors

Kuncheng Song[1], Hayden N Brochu[1], Monica L Bustos[1], Qimin Zhang[1], Crystal R Icenhour[1], Stanley Letovsky[1] and Lakshmanan K Iyer[1]
[1]Labcorp, Burlington, NC 27215, USA

Abstract

The nasopharyngeal microbiome acts as a dynamic interface between the human body and environmental exposures, modulating immune responses and helping to maintain respiratory health. Despite its clinical significance, the nasopharyngeal microbiome remains poorly characterized with few resources available. A robust standardized framework for distinguishing healthy from dysbiotic communities is needed to advance nasopharyngeal microbiome research.

We conducted a comprehensive meta-analysis that yielded 28 independent nasopharyngeal microbiome studies encompassing 7,790 uniformly reprocessed samples from public repositories. Further, we developed a newly introduced, three-stage large language model-assisted background decontamination pipeline that mitigates study-specific technical and environmental variation while preserving biological signal integrity.

We identified six reproducible nasopharyngeal community state types (NPCSTs), each dominated by distinct genera: *Moraxella* (I), *Corynebacterium-Dolosigranulum* (II), *Staphylococcus* (III), *Haemophilus* (IV), mixed (V), and *Streptococcus* (VI). NPCSTs were key drivers of the sample diversity rather than disease status, requiring NPCST-stratified downstream analyses. Co-occurrence network analysis and functional pathway predictions revealed NPCST-specific microbial interactions and metabolic signatures.

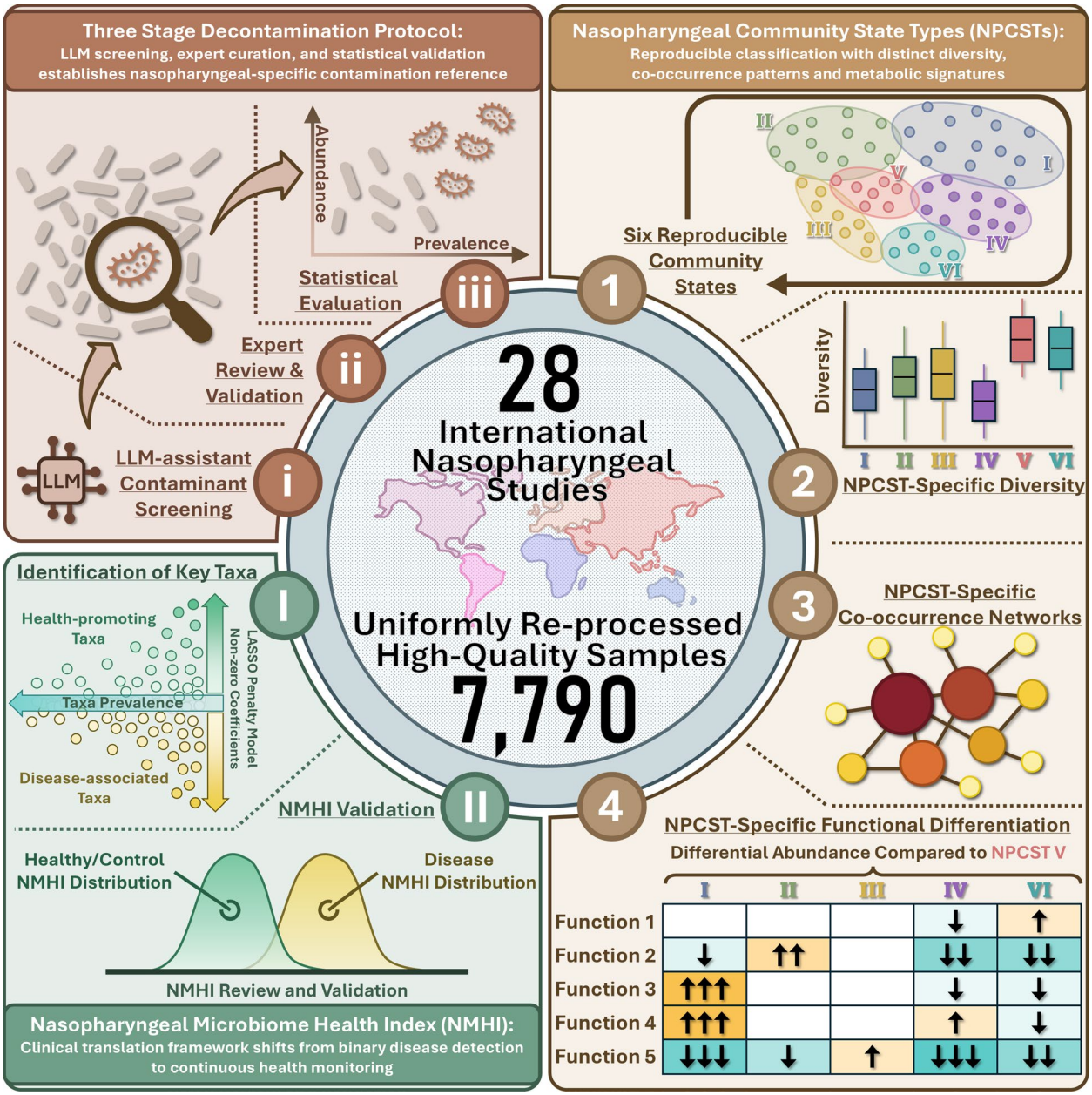
To bridge the translational gap from research to clinic, we developed the Nasopharyngeal Microbiome Health Index (NMHI), an independent continuous wellness score that achieves an AUC of 0.90 and 0.92 in internal and external validations, respectively. Unlike binary disease classifiers, the NMHI quantifies nasopharyngeal health along a spectrum, enabling risk stratification for respiratory infections, treatment monitoring, and dysbiosis identification in PCR-negative symptomatic patients.

Our comprehensive framework establishes reproducible NPCSTs, validated decontamination protocols, and publicly available classification models, providing the analytical framework for advancing nasopharyngeal microbiome research toward clinical implementation. The NMHI exemplifies international best practices for microbiome diagnostics through rigorous external validation, standardized pipelines, and clinically interpretable wellness metrics. It further delivers an evidence-based framework for translating nasopharyngeal microbiome research into precision respiratory medicine, enabling future development of microbiome-based respiratory therapeutics.

Keywords

Community State Type; Microbiome; 16S; Meta-Analysis; Co-occurrence Network; Functional Prediction; Machine Learning

Graphic Abstract



Highlights

- We identified six reproducible nasopharyngeal community state types (NPCSTs) that transcend study-specific variation, providing a standardized framework for classifying nasopharyngeal microbiome structures across diverse populations.
- Our newly introduced three-stage decontamination pipeline combining LLM-assisted screening with expert curation successfully eliminates environmental artifacts in low-biomass nasopharyngeal samples while preserving biological signal integrity.
- Diversity metrics vary across NPCSTs independent of disease status, while co-occurrence networks and functional profiles show NPCST-specific patterns, collectively requiring community-stratified analyses to prevent misleading health-disease interpretations.
- The Nasopharyngeal Microbiome Health Index (NMHI) shifts respiratory assessment from binary pathogen detection to continuous microbiome wellness monitoring, identifying individuals with dysbiotic nasopharyngeal microbiomes.

Introduction

The nasopharynx, positioned at the convergence of the respiratory and digestive tracts, serves as a critical reservoir of microbial colonization, where its anatomically protected mucosa harbors higher pathogen densities while minimizing contamination compared to oral or nasal surfaces[1]. Established early as the optimal site for recovering viable respiratory pathogens[2], it became the gold-standard location for respiratory diagnostics[3], underpinning PCR-based panels including the BioFire® Respiratory Panel[4] which targets 22 pathogens and Labcorp seasonal panels[5]. Each year, millions of nasopharyngeal swabs are collected in the United States for diagnostic tests including RSV[6] and others, with volumes amplified during the COVID-19 pandemic to 17.4 million tests per week[7]. With even higher test volumes globally, PCR-based testing of nasopharyngeal samples plays a critical role in respiratory health surveillance. However, systematic investigation of the nasopharyngeal microbiome remains surprisingly limited. This knowledge gap became critically apparent during the COVID-19 pandemic, when conflicting studies reported either significant microbiome differences[8–10] or no alterations[11] between disease severities, exposing our fundamental inability to interpret how resident nasopharyngeal microbial communities shape disease trajectories.

Four critical limitations have prevented characterization of the core nasopharyngeal microbiome and realization of its diagnostic and therapeutic potential. First, the absence of standardized protocols for removing environmental and reagent-derived

contamination[12,13], particularly problematic in low-biomass nasopharyngeal samples[14], renders both individual studies and meta-analyses vulnerable to spurious signals that obscure true biological patterns. Existing decontamination methods (e.g., decontam[15], micRoclean[16] and CleanSeqU[17]) cannot be applied to meta-analyses, as negative control samples are typically unavailable. Moreover, even with negative controls, these methods cannot effectively identify sample-specific environmental contaminants, which become disproportionately enriched in low-biomass samples and can dominate the microbial signal.

Second, attempts to define distinct nasopharyngeal community structures have yielded study-specific clusters that fail to replicate across cohorts. The nasopharyngeal microbiome, like other mucosal sites dominated by specific genera (e.g., vaginal *Lactobacillus*[18,19]), exhibits pronounced dominance patterns[18,19], making community-level analysis crucial as these dominant taxa carry biological importance. Current studies report widely varying numbers of clusters (4 to 13)[20–22], with some approaches using unsupervised clustering of all detected taxa while others pre-select known nasopharyngeal genera to guide analyses[10,23]. However, no existing framework has demonstrated stability and reproducibility of nasopharyngeal community state types (NPCSTs) across independent datasets.

Third, machine learning approaches for microbiome-based prediction demonstrate unstable performance, with models rarely undergoing external validation and identified biomarker taxa proving non-generalizable across populations. Most studies rely solely on differential abundance tests such as ANCOM-BC[24], MaAsLin2[25], or sparse compositional modeling[22] to identify differentially abundant phyla[24], genera[24,26,27], species[28], amplicon sequence variants (ASVs)[29–31], or operational taxonomic units (OTUs)[22,32]. However, these univariate approaches cannot advance nasopharyngeal understanding from research to clinical practice, as they produce study-specific markers that fail to replicate across cohorts. Even studies employing machine learning models achieve low to modest performance (AUC 0.59-0.82)[9,26,33–35], with this variability reflecting overfitting to individual datasets rather than capturing generalizable patterns. These limitations underscore the critical need for externally validated, reproducible models to enable clinical translation of nasopharyngeal microbiome insights.

Fourth, clinical translation remains limited not by competition with gold-standard qPCR diagnostics, but by an incomplete understanding of the nasopharyngeal microbiome's complementary diagnostic value. Current approaches lack tools for monitoring qPCR-negative cases, including suspected infections with negative results, where microbiome analysis could reveal underlying dysbiosis that informs clinical outcomes and disease

severity[36]. This gap is particularly critical for symptomatic patients with negative pathogen tests and individuals at high risk for severe outcomes who could benefit from therapeutic interventions.

To address these fundamental limitations, we conducted a comprehensive meta-analysis of nasopharyngeal microbiome studies and developed a unified analytical pipeline that systematically addresses each methodological challenge. We implemented a newly introduced three-stage decontamination strategy integrating large language model-assisted screening, expert curation, and statistical validation to establish a comprehensive nasopharyngeal-specific contamination reference database. We also characterized six reproducible nasopharyngeal community state types (NPCSTs) each with unique microbial signatures. We then characterized the distinct and shared features of these NPCSTs by analyzing their microbial co-occurrence networks and predicted functions.

To improve clinical interpretability, we developed and validated the Nasopharyngeal Microbiome Health Index (NMHI), which quantifies respiratory health status directly from microbiome composition. We rigorously validated each NMHI component using external, independent datasets to ensure precision, robustness and generalizability across all NPCSTs.

Our reproducible framework establishes the foundation for nasopharyngeal microbiome health assessment and creates a clear pathway for clinical translation. We address multiple methodological challenges and provide the necessary resources to advance respiratory microbiome research from exploratory studies to clinical implementation.

Methods

Study screening and metadata evaluation

Our meta-analysis followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (see **Supplementary File** for PRISMA Checklist)[37]. We systematically searched PubMed[38] and Embase[39] on May 14th, 2024, using the query: ("nasopharyngeal") AND ("microbiome" OR "microbiota"). Two independent reviewers applied the exclusion criteria detailed in **Figure 1B**. Initial title and abstract screening of 710 publications yielded 227 for full-text review. Among these, we identified 188 studies with complete metadata (age, disease, location, publication year) regarding nasopharyngeal sample usage. We then aggregated the disease and geographic data by research topic (**Figure 1C**), counting studies addressing multiple topics in their respective categories before consolidating into 10 higher-level categories. Of these 227 studies, only 63 provided downloadable nasopharyngeal swab 16S rRNA gene sequencing data. After

applying our data processing and background decontamination protocols (see subsequent **Methods**), we retained 28 studies comprising 7,790 samples for final analysis. This meta-analysis curation process was also included in our independent study by Bustos et al. that details inconsistencies and the need for data standards in nasopharyngeal research.

16S data processing

Raw 16S rRNA gene sequencing datasets were obtained from NCBI BioProjects listed in **Table 1**. All raw reads were processed using a standardized bioinformatics pipeline to identify hypervariable regions (HVRs), perform quality control, and generate amplicon sequence variants (ASVs) (see **Supplementary File** for detailed processing parameters and quality control criteria). Briefly, automated HVR identification and confirmation was performed using kmer hashing[40] and primer detection (**Tables S1** and **S2**), followed by quality filtering, trimming, and denoising using DADA2 v1.22.0[41]. After removing singletons and chimeras, ASVs were taxonomically classified using the SILVA v138.2 database[42] and DADA2-implemented RDP classifier at a bootstrap confidence threshold of 80. ASVs lacking family-level classification were excluded, and sequences were aggregated at the lowest assigned taxonomic rank to generate count matrices for each study. To ensure sufficient sequencing depth for reliable diversity estimates, samples with fewer than 5,000 family-level reads following background decontamination were excluded (see subsequent **Methods** section). Studies with >50% of samples passing all quality filters (n=28) were retained to ensure dataset-wide consistency for meta-analysis.

Nasopharyngeal background decontamination protocol

We developed and validated a newly introduced decontamination strategy to eliminate environmental- and laboratory-derived background contaminant signals from nasopharyngeal microbiome samples, including reagent-associated taxa (i.e., the "kitome")[13,43], marine/aquatic organisms, and soil/sediment-derived bacteria that represent technical artifacts rather than authentic nasopharyngeal colonizers. Our three-stage approach integrated: (1) LLM-assisted (Claude[44]) contamination screen (in triplicate) with a standardized contamination assessment prompt (**Supplementary File**), (2) human expert review of triplicate results and generation of contaminant bacterial family catalog, and (3) statistical validation of flagged taxa prevalence to confirm their sparse, study-specific distribution across datasets (**Figure 2A**). Independent triplicate evaluation using Claude[44] with a standardized contamination assessment prompt (**Supplementary File**), followed by expert harmonization. This pipeline generated a nasopharyngeal-specific contamination reference database for future research applications. To ensure reproducibility and computational efficiency while managing thousands of genera and species, we performed contamination assessment at the family taxonomic level. We then

systematically removed all genera, species, and ASVs belonging to flagged families, generating study-specific decontaminated count matrices. Rigorous validation of our pipeline using 28 studies and two independent validation cohorts confirmed preservation of biological signal integrity (**Supplementary File**).

Phylogenetic analysis

Since the meta-analysis spanned multiple HVRs (**Table 1**), the representation of various taxa across HVRs was assessed (**Figure S1**). Based on these findings, V3-V4 ASVs were selected for phylogenetic analysis. Because sequencing errors are known to propagate into ASVs and exhibit study-specific patterns, ASVs were clustered and filtered while preserving taxonomic representation of the full dataset. ASVs were initially clustered using CD-HIT v4.8.1[45] with parameter -c 0.99. Representative ASVs were subsequently selected from each cluster by prioritizing those with unique and highly resolved taxonomic assignments. For example, if two ASVs from genus A clustered with another ASV from genus A resolved to species A, only the species-level ASV was retained. All distinct taxa within each cluster were preserved; thus, an ASV from genus B in the same cluster was also included. For ties (e.g., two ASVs from genus A species A), ASVs were selected based on study-wise prevalence (number of study detections), with remaining ties resolved by sample-wise prevalence (total number of sample detections). Following this clustering process, ASVs underwent final screening that required detection in a minimum of two studies, yielding a parsimonious set of representative V3-V4 ASVs.

Phylogenetic analysis was then performed by generating a multiple sequence alignment of all representative V3-V4 ASVs using Clustal Omega v1.2.4[46] and with subsequent processing in R[47] (v.4.2.2) using the phangorn v2.8.1[48] package. A neighboring joining tree was constructed using the dist.ml and NJ functions. Then, a Jukes–Cantor optimized maximum likelihood tree was generated using the pml and optim.pml functions with optNni = T. The ggtree v3.2.0[49] R package was used to visualize the resulting phylogeny.

Nasopharyngeal Community State Types (NPCST) classification

Since genus-level resolution was more stable that species-level resolution (**Figure S2**), a minimum relative abundance (RA) threshold of 0.01% was applied to remove rare taxa at the most refined resolution across all samples before aggregation to the genus level for consistent analyses. Taxa that did not resolve a genus (i.e., only family-level resolution) were retained and denoted with the prefix ‘f_’, while genera were denoted with ‘g_’. This filtered dataset was utilized for this and all subsequent analyses. Nasopharyngeal microbiome profiles were hierarchically clustered using Ward’s minimum variance method with the pairwise Bray-Curtis dissimilarity matrix computed using the vegdist function in

vegan v2.6–8[50]. Hierarchical cluster selection was performed via cluster v2.1.6[51] silhouette score analysis, with the optimal tree cut determined by selecting the highest silhouette score that simultaneously maximized the number of clusters.

The Unsupervised Learning based Definition of the Rare Biosphere (ulrb)[52] method was employed to classify abundant, undetermined/intermediate, and rare taxa across different NPCSTs. This approach operates independently of RA thresholds and provides more reliable identification of key taxa that define these 6 NPCSTs. To validate the robustness and reproducibility of the NPCSTs, leave-one-study-out (LOSO) cross-validation (CV) was implemented, wherein each study was systematically removed and the remaining 27 studies were used to regenerate the NPCSTs using the aforementioned methods. The top 2-20 clusters were subsequently identified and cluster membership consistency was evaluated. Additionally, two machine learning classifiers (SVM and Random Forest) were validated and made publicly available for NPCST assignment in future studies. Detailed methodology is provided in the **Supplementary File**.

To validate the statistical significance of the identified clusters, permutational multivariate analysis of variance (PERMANOVA) was performed using the vegan adonis2 function to test for overall differences in microbial community composition among all NPCSTs, with age, hypervariable regions (HVR), sex, and study included as covariates. Analyses incorporating age and sex were performed on a subset of samples due to limited metadata availability.

Validation dataset

To validate the analytical approach, two additional BioProjects were selected from publications released after the initial data extraction cutoff. The validation cohort comprised: (i) PRJNA981220, encompassing SARS-CoV-2 samples (n=40); and (ii) PRJNA914884, containing respiratory syncytial virus (RSV) samples (n=1,537). These datasets were processed using identical methodological protocols as previously described, resulting in a final validation dataset of 1,577 samples. NPCST classifications for validation samples were assigned using a random forest model trained on the meta-analysis samples and their corresponding NPCST assignments (**Supplementary File**). The external datasets for the nasopharyngeal microbiome health index (NMHI) used a combination of longitudinal and cross-sectional samples that were independent and were not included in the training model. Further, samples with collection date outside the symptomatic period and those without clear healthy/disease classifications (e.g., infants less than 1 month old) were also excluded from it.

Co-occurrence network analysis

SPIEC-EASI (v1.1.2)[53] was employed to infer microbial associations both globally and within individual NPCSTs with the 'mb' algorithm applied to genus-aggregated abundance matrices containing both family- and genus-resolved taxa. Only cross-sectional samples (n=5,816 of 7,790 total) were analyzed to prevent bias from repeated measures. Separate microbial interaction networks were constructed for the complete cross-sectional dataset and each NPCST subset. Association directionality was determined from edge signs (<0 for mutual exclusion, >0 for co-occurrence) after adjusting for confounding taxon-taxon relationships. Network optimization criteria included ≥1% prevalence and >10 read thresholds, retaining 264-525 taxa per NPCST from 626 total taxa. Across all NPCST-specific networks, 72 taxa exhibited detectable associations. Hub taxa were identified through k-core decomposition with igraph (v2.1.1)[54], with the upper quintile (top 20%) of k-core values defining biologically relevant and structurally important network components.

Prediction of metabolic pathways

Functional profiling was performed using PICRUSt2-SC[55] with default parameters to predict MetaCyc[56] Enzyme Commission (EC) metabolic pathways from 16S rRNA ASV sequences and abundances, restricted to ASVs assigned to the family- and genus-resolved taxa in the final genus-aggregated abundance matrix (see 'NPCST classification' **Methods** section). To identify differentially abundant pathways across NPCSTs, both MaAsLin3[57] and LinDA[58] were applied to PICRUSt2-SC predicted pathway abundances from all cross-sectional samples (n=5,816, see 'Co-occurrence network analysis' **Methods** section). Both methods employed identical fixed-effects models with NPCST (using NPCST V as the reference) and BioProject as covariates, 10% prevalence filtering, and Benjamini-Hochberg correction ($q < 0.05$). MaAsLin3 implemented total sum scaling normalization, log transformation, and median comparison for compositional data, while LinDA utilized adaptive winsorization (97th percentile) within its compositional framework. Consensus pathways were defined as those achieving significance using both methods with concordant directional effects.

For NPCST-specific characterization, asymmetric thresholds were applied due to unequal distributions of enriched versus depleted pathways within each NPCST. To capture equivalent numbers of the most significant positive and negative effects, we implemented NPCST-specific asymmetric thresholds to identify high-impact functional signatures. Thirty-eight pathways were prioritized based on consensus significance, effect magnitude, and metabolic category representation, and were classified as either shared (multi-NPCST) or unique (NPCST-specific). This dual-method framework distinguished between changes in functional potential (pathway presence/absence via prevalence models) and

metabolic activity (pathway abundance when present). Thirty-eight pathways were prioritized based on consensus significance, effect magnitude, and metabolic category representation, and were classified as either shared (multi-NPCST) or unique (NPCST-specific). This dual-method framework distinguished between changes in functional potential (pathway presence/absence via prevalence models) and metabolic activity (pathway abundance when present), revealing NPCST-specific ecological strategies.

Nasopharyngeal Microbiome Health Index Development

Building upon the Gut Microbiome Wellness Index 2 (GMWI2) framework[59] as a conceptual foundation, we developed the Nasopharyngeal Microbiome Health Index (NMHI) through substantial methodological innovations specifically tailored to address the sparse, highly variable nasopharyngeal microbiome characterized by distinct community structures. We applied LASSO-penalized logistic regression with balanced class weights ($w_{healthy} = 0.5/(n_{healthy}/n_{total})$, $w_{disease} = 0.5/(n_{disease}/n_{total})$) to all 5,435 cross-sectional samples with interpretable disease/healthy status (see 'Validation dataset' **Methods** section), using binary presence/absence encoding at a 0.01% RA threshold. Given that NPCSTs explained substantially more variance than study effects (53.19% vs. 13.06%), we implemented Leave-One-NPCST-Out (LONO) CV for the initial lambda selection across seven taxonomic configurations (phylum through species, and all levels combined).

We created binary classifications distinguishing healthy controls from either viral infections alone or all conditions combined. Based on comprehensive validation, we focused on four models combining condition scope (all conditions vs. viral infections only) and taxonomic resolution (all-taxa vs. genus-only).

Our five-stage methodology comprised: **(1)** LONO CV for lambda selection (0.0001-0.03 range), maximizing AUC across held-out NPCSTs; **(2)** performance evaluation using LONO CV and 10-fold CV across five prevalence thresholds (0-20%), selecting 5% for optimal performance-interpretability balance; **(3)** final model training to extract coefficients, with NMHI calculated as intercept plus sum of present taxa coefficients; **(4)** threshold optimization (-5 to 5, 0.1 increments) maximizing balanced accuracy for global and NPCST-specific cutoffs; **(5)** external validation on 699 independent samples: healthy family members with recurrent respiratory infections (n=265), varying SARS-CoV-2 severity (n=398 from different studies), lower respiratory tract infections (n=5), and non-SARS-CoV-2 critically ill patients (n=31) (see 'Validation dataset' **Methods** section). Model performance was evaluated using AUC and balanced accuracy metrics. Complete methodology and results are provided in the **Supplementary File**.

Statistical methods

All statistical analyses were performed in R[47] v4.2.2. Alpha diversity metrics were calculated after rarefying samples to 5,000 reads to standardize for sequencing depth variation (range: 5,009–646,618; median: 30,204 reads), computing genus richness, Shannon diversity, and Inverse Simpson indices using the vegan[50] package. Beta diversity and clustering approaches, including Bray-Curtis dissimilarity calculations, hierarchical clustering, and PERMANOVA analyses, are detailed in their respective **Methods** sections above.

To validate the preservation of biological signals in our decontamination pipeline, we employed Procrustes analysis to assess ordination congruence (correlation coefficient and M^2 statistic) and Mantel tests to evaluate distance matrix correlations between before and after decontamination datasets, with significance determined using 999 permutations (vegan[50] package).

NPCST associations with host demographics and disease status were evaluated using 5,614 samples with complete metadata (5,943 additional samples contributed to sex-only analyses). We compared age distributions across NPCSTs using Kruskal-Wallis tests with η^2 effect sizes (rstatix[60]) and sex proportions via χ^2 tests of independence with Cramér's V effect sizes (DescTools[61]). Pairwise comparisons employed Wilcoxon rank-sum tests with Benjamini-Hochberg correction. To quantify disease susceptibility by NPCST, we fitted conditional multinomial logistic regression models (mclogit[62] v0.9.6) with NPCST membership as outcome and infection type (bacterial, viral, or control) as predictor, stratified by BioProject. Marginal probability contrasts were calculated using emmeans[63] (v1.10.5) with FDR adjustment, generating odds ratios with 95% confidence intervals.

Model performance was assessed through multiple metrics: area under the receiver operating characteristic curve (AUC) with 95% confidence intervals, balanced accuracy at optimized thresholds, and Cohen's d effect sizes for group discrimination. For functional pathway concordance between MaAsLin3[57] and LinDA[58] methods, we calculated concordance correlation coefficients (CCC), Pearson correlation, and Spearman's rank correlation. All group comparisons utilized Wilcoxon rank-sum tests with Benjamini-Hochberg correction for multiple testing adjustment.

Results

Literature Screening and Study Selection Process

Following our systematic review protocol (**Methods**), we analyzed 28 studies comprising 7,790 samples each with paired 16S rRNA gene sequencing data, with study characteristics detailed in **Table 1** and **Figure 1**. For external validation, we subsequently identified 2 additional studies containing 1,573 samples using identical selection criteria.

Table 1. Meta-analysis study characteristics. Each row shows the details of one of the 28 studies, including the data source (NCBI BioProject accession), hypervariable region (HVR), reference, samples analyzed and passing quality control (QC), population characteristics (age range), and disease investigated.

Disease/Condition of Interest	Reference	Age Range (Years)†	PMID	NCBI BioProject	Samples Available for Download	Hypervariable Region	Samples Passing QC	Samples Passing QC [%]	Cohort Type
Respiratory viruses	Edouard et al., 2018[64]	0-94	30033505	PRJEB14780	451	V3-V4	307	68.07%	Main
Chronic Rhinosinusitis	De Boeck et al., 2019[65]	18-65	31776238	PRJEB30316	172	V4	151	87.79%	Main
Otitis Media	Jörissen et al., 2021[66]	~0 - 18	33879499	PRJEB33591	114	V4	113	99.12%	Main
Cystic Fibrosis	Pittet et al., 2021[67]	<18	33247928	PRJEB41059	5	V3-V4	5	100.00%	Main
Pneumococcal Colonization on HIV exposed/unexposed	Kelly et al., 2022[25]	<1	34511605	PRJEB41741	156	V3-V4	121	77.56%	Main
Influenza-like Illness	Shetty et al., 2022[68]	~63.5-77.6	35115596	PRJEB46215	1,072	V4	878	81.90%	Main
Asthma	Thorsen et al., 2023[69]	<1	37863895	PRJEB58215	290	V3-V4	268	92.41%	Main
Meningococcemia	Bozan et al., 2023[70]	<3	37370879	PRJEB59932	49	V3-V4	25	51.02%	Main
Tuberculosis	Ruiz-Tagle et al., 2023[71]	~32-39	37147354	PRJEB61078	82	V1-V3	24	29.27%	Main
Otitis Media	Lappan et al., 2018[22]	0.108 - 0.267	29458340	PRJNA388707	195	V3-V4	181	92.82%	Main
HIV and Tuberculosis	Htun et al., 2018[72]	Not available	30364728	PRJNA432583	33	V3-V4	31	93.94%	Main
Enterovirus D68	Bowers et al., 2019[73]	1.5-17	30670612	PRJNA474932	6	V3-V4	4	66.67%	Main
SARS-CoV-2	Ventero et al., 2021[74]	13-94	33815323	PRJNA673585	74	V3-V4	74	100.00%	Main
SARS-CoV-2	Merenstein et al., 2021[26]	36-94	34399607	PRJNA683617	267	V1-V2	245	91.76%	Main
SARS-CoV-2	Braun et al., 2021[11]	30-68	33903709	PRJNA688646	55	V4	30	54.55%	Main
SARS-CoV-2	Hurst et al., 2022[31]	4-17	35247047	PRJNA703574	285	V4	260	91.23%	Main
Exacerbated Wheezers and Asthmatics	Aydin et al., 2021[32]	0.25 - 17	35386977	PRJNA714100	64	V3-V4	64	100.00%	Main
Respiratory Tract Infection	Piters et al., 2022[30]	0-1	35058634	PRJNA740120	1,156	V4	1,142	98.79%	Main
SARS-CoV-2	Galeeva et al., 2022[75]	Not available	34977286	PRJNA751478	335	V3-V4	229	68.36%	Main
SARS-CoV-2	Ventero et al., 2022[36]	52-80	34963638	PRJNA754005	177	V3-V4	171	96.61%	Main
SARS-CoV-2	Crovetto et al., 2022[24]	27.3-35.8	35927569	PRJNA777915	76	V3-V4	75	98.68%	Main
Environmental Impact	Karampatsas et al., 2023[76]	20-69	37525095	PRJNA788869	44	V4	41	93.18%	Main
SARS-CoV-2	Ferrari et al., 2022[77]	~33-57	35873175	PRJNA839581	54	V3-V4	51	94.44%	Main

Lower Respiratory Tract Infection	Koenen et al., 2023[78]	2-6	37950996	PRJNA897835	358	V4	265	74.02%	Main
SARS-CoV-2	Kim et al., 2023[27]	37-77	36583481	PRJNA912320	66	V1-V2	58	87.88%	Main
Pneumococci Carrier	Paulo et al., 2023[29]	25-50	37658443	PRJNA962709	177	V4	152	85.88%	Main
SARS-CoV-2	Romani et al., 2024[79]	<17.7	38289047	PRJNA993438	106	V3-V4	99	93.40%	Main
Healthy	Odendaal et al., 2024[80]	0-87	39094567	PRJNA997934	2,799	V4	2,726	97.39%	Main
SARS-CoV-2	Galeana-Cadena et al., 2024[75]	20-87	38826746	PRJNA981220	40	V4	40	100.00%	External
RSV	Kristensen et al., 2024[34]	< 1	39642873	PRJNA914884	1,537	V4	1,533	99.74%	External

†The symbol "~" indicates that exact ranges were not provided in the original source; values represent approximations based on reported means ± standard deviations or other available metrics.

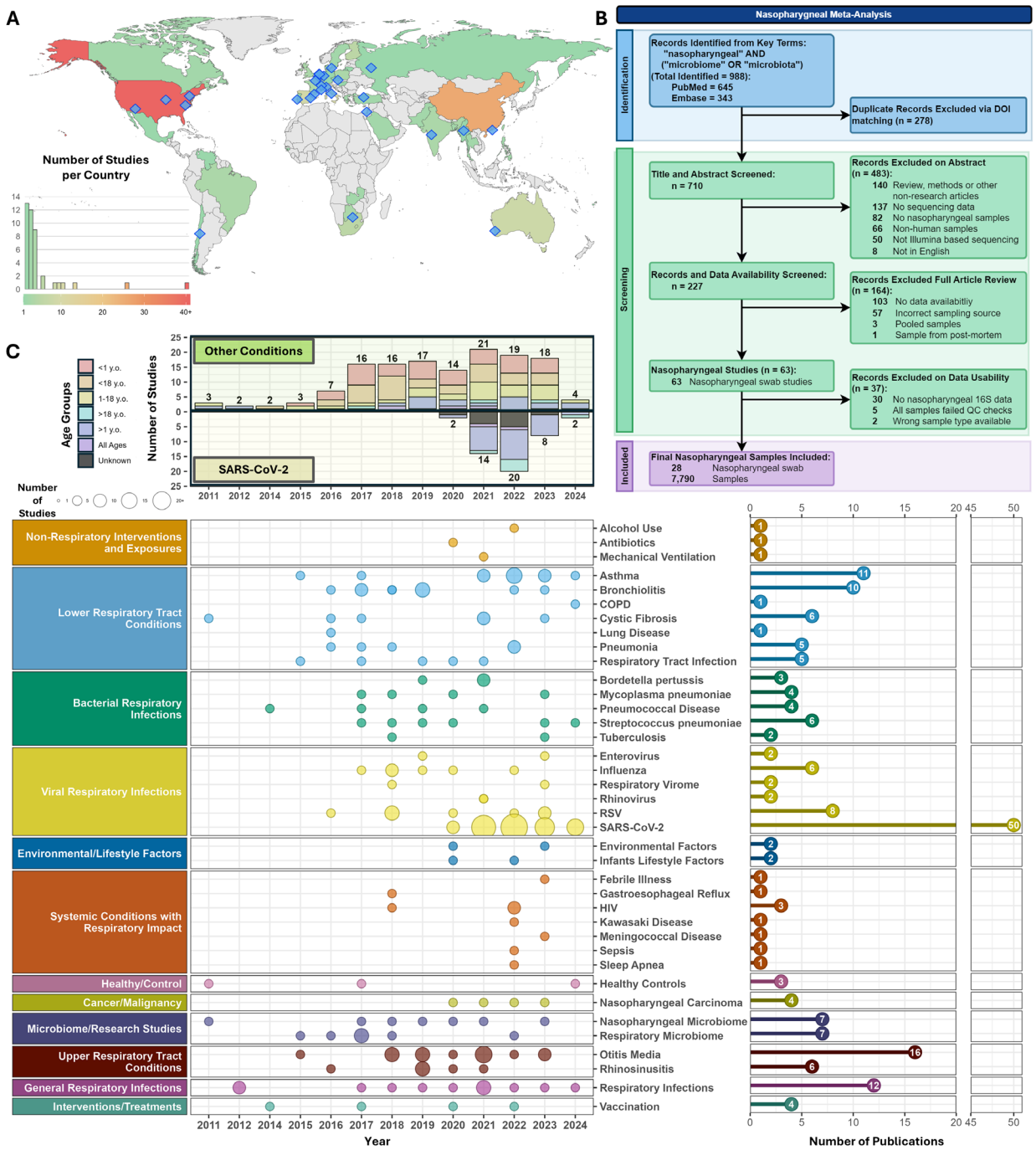


Figure 1. Global distribution and temporal trends of nasopharyngeal microbiome studies. A. World map displaying geographic distribution of publications examining human nasopharyngeal samples (n=181 across 42 countries). Color intensity indicates publication frequency per country (green to red gradient, 0-40+ studies). Blue diamonds mark locations of the 28 studies with downloadable datasets. **B.** PRISMA flow diagram for the systematic review. **C.** Temporal distribution of nasopharyngeal microbiome research by topic and age category (2011-2024). Studies are categorized into 10 high-level research topics encompassing 39 specific conditions, with bubble size proportional to annual study count and color-coded by research topic. Studies addressing multiple topics are counted

in each relevant category. Upper panel: stacked bar plot depicting distribution of study publication years with colors distinguishing studies by age cohort studied. Studies are separately counted by condition (bottom: SARS-CoV-2 infection, top: other conditions). Right panel: total number of publications per condition.

Global and temporal trends of nasopharyngeal microbiome studies

From the 227 full-text reviewed articles, we identified 181 studies that directly examined nasopharyngeal samples (including both swab and aspirate samples) across 42 countries. The United States and China led research output with 41 and 26 studies, respectively, while most of the other research activity was broadly dispersed across European countries (**Figure 1A**). Nasopharyngeal microbiome research emerged in 2011 and accelerated significantly after 2017. Pre-2021 studies predominantly examined pediatric populations (infants <1 year; children 1-18 years), reflecting the clinical relevance of nasopharyngeal sampling in pediatric respiratory health. The COVID-19 pandemic catalyzed a pivotal shift: adult-focused studies increased substantially post-2021, driven by SARS-CoV-2 investigations that leveraged nasopharyngeal sampling for concurrent viral detection and microbiome analysis. During this shift, non-COVID studies maintained their pediatric focus.

Our meta-analysis encompassed a broad spectrum of research topics, including respiratory tract infections of both bacterial and viral origin across diverse disease conditions. Bacterial infections included chronic rhinosinusitis[65], meningococcemia[70], otitis media[22,66], pneumococcal colonization[25], pneumococcal carriage[29,78], and tuberculosis (including HIV-TB co-infections)[71,72]. Viral infections comprised enterovirus D68[73], influenza-like illness[68], multiple respiratory viruses[30,64], and numerous SARS-CoV-2 studies[11,24,26,27,31,36,74,75,77,79]. We also included lower respiratory conditions and chronic inflammatory diseases, such as asthma[32,69] and cystic fibrosis[67]. Additional studies examined environmental impacts[76] and healthy populations[80] for comprehensive comparative analyses (**Figure 1C**). Despite this breadth, significant gaps persist in understanding how the environment, antibiotics, and/or vaccination are associated with nasopharyngeal microbiome community structure (**Figure 1C**).

Three-Stage Decontamination Pipeline

Approximately 477M paired-end 16S rRNA sequencing reads from 8,699 samples were uniformly processed to yield approximately 331M family-resolved amplicon sequence variant (ASV) reads (~69.5% of raw input). We then required a sample minimum of 5,000 family-resolved ASV reads, yielding 8,314 samples with 330M reads representing 596 bacterial families. Due to the absence of negative control and diversity of lab protocols used in the studies, we devised a three-stage, LLM-assisted (Claude[44]) decontamination

pipeline to remove batch effects and study-specific contamination patterns (**Figure 2A**). We screened 596 unique family-level taxa and identified 532 families (89.3% of 596) as environmental or technical contaminants, with soil, aquatic, anaerobic, and thermal environments as primary contamination sources (**Table S3**). These contaminants exhibited significantly lower prevalence (median: 0.2% vs. 15.4% for authentic signal; Wilcoxon rank-sum test, $p < 2.2 \times 10^{-26}$) and 3.5-fold lower median relative abundance (RA) (0.070% vs. 0.244%; Wilcoxon rank-sum test, $p < 2.2 \times 10^{-26}$) compared to authentic nasopharyngeal taxa (**Figure 1B**).

Prevalent and abundant contaminant families broadly came from the following sources (**Figure 2B**, 11 families detected in >25% of samples with median RA >0.1%). The three most prevalent families were common water and reagent system inhabitants *Comamonadaceae*[13], *Pseudomonadaceae*[13], and *Burkholderiaceae*[13], while others originated from soil and environmental sources: *Rhizobiaceae*[13], *Sphingomonadaceae*[13], *Beijerinckiaceae*[13], *Xanthobacteraceae*[81], *Thermaceae*[82], *Caulobacteraceae*[13], *Chitinophagaceae*[13] and *Nocardiaceae*[83] (**Figure 2B**). After removal of all contaminant families (~12% of family-resolved ASV reads), we restricted the dataset to samples with at least 5,000 authentic family-resolved ASV reads (n=7,790; 30.2K median reads). This process maintained high sample retention across the 28 studies (**Figure S1**, 84.23% median), while ensuring sufficient sequencing depth for diversity analyses (**Table 1** and **Figure S2A-B**). Intriguingly, decontamination of our external validation datasets identified only three novel genera at negligible abundances (~0.02%) and no additional family-level contaminants (**Supplementary File**). Together, these findings demonstrate consistent decontamination and comprehensive recovery of the nasopharyngeal microbiome across diverse populations, confirming pipeline applicability to new cohorts.

We then compared genus-resolved microbiome profiles from before and after decontamination, and we investigated if/how decontamination influenced our nasopharyngeal microbiome community state types (NPCSTs) analysis described later (**Figure S3**). We found that 81.2% of samples (6,327 of 7,790) maintained their original NPCST classification following decontamination (**Figure S3A-B**). Procrustes analysis demonstrated near-perfect preservation of sample relationships (correlation = 0.973, M^2 = 0.05, $p < 0.001$), corroborated by Mantel testing showing strong correlation between before and after decontamination distance matrices ($r = 0.971$, $p < 0.001$). Study-associated batch effects were reduced (**Figure S3C-D**, PERMANOVA R^2: 13.97% to 13.06%), and 1,810 ultra-low prevalence contaminating genera (74.3% of 2,436) were removed from the dataset, yielding a 2.36% sparsity reduction and substantially improved statistical power for downstream analyses. These convergent validation metrics confirm that our

decontamination protocol removes environmental artifacts while preserving authentic nasopharyngeal microbiome structure.

Taxonomic characteristics of nasopharyngeal microbiomes

We began taxonomic analysis by investigating the resolution of 16S hypervariable regions (HVRs) targeted by studies in this meta-analysis. Most studies (25 of 28, 89%) targeted V3-V4 or V4 hypervariable regions for 16S rRNA gene sequencing (**Table 1**), with similar quantification of genus-resolved taxa between the V3-V4 and V4 HVRs (**Figure S2C**). We observed better phylogenetic resolution in the V3-V4 region, as indicated by a higher proportion of genus- and species-resolved reads among studies targeting the V3-V4 HVR compared to the V4 HVR (**Figure S2C-D**). Both HVRs exhibited reduced species-level classification reliability, failing to resolve a species for 83.3% of reads (sample median, IQR: 68.0-94.3%). However, substantially more reads resolved a genus (median: 99.4%, IQR: 97.6-99.9%) (**Figure S2C-D**), indicating that both the V3-V4 and V4 HVRs consistently resolved genus-level diversity, ensuring reliable nasopharyngeal community structure analysis.

Phylogenetic analysis of 1,091 V3-V4 ASVs representing the 64 authentic bacterial families (**Methods**) revealed 14 core families that were highly prevalent (**Figure 2C**, detected in >50% of 7,790). *Streptococcaceae* was the most prominent family, colonizing 95.46% of samples, followed by *Corynebacteriaceae*, *Moraxellaceae*, *Staphylococcaceae*, *Carnobacteriaceae*, and *Pasteurellaceae* (76.57%-93.13%; **Figure 2C-D**). These 14 core families harbored a median of 12.5 unique genera each, with consistently elevated prevalence relative to the remaining 50 families (**Figure 2C**, outer heatmap of the phylogeny). Collectively, these 14 families comprised a median 97.1% (IQR: 91.4-99.5%) of total microbial abundance, establishing the fundamental taxonomic architecture of the nasopharyngeal microbiome (**Figure 2D**).

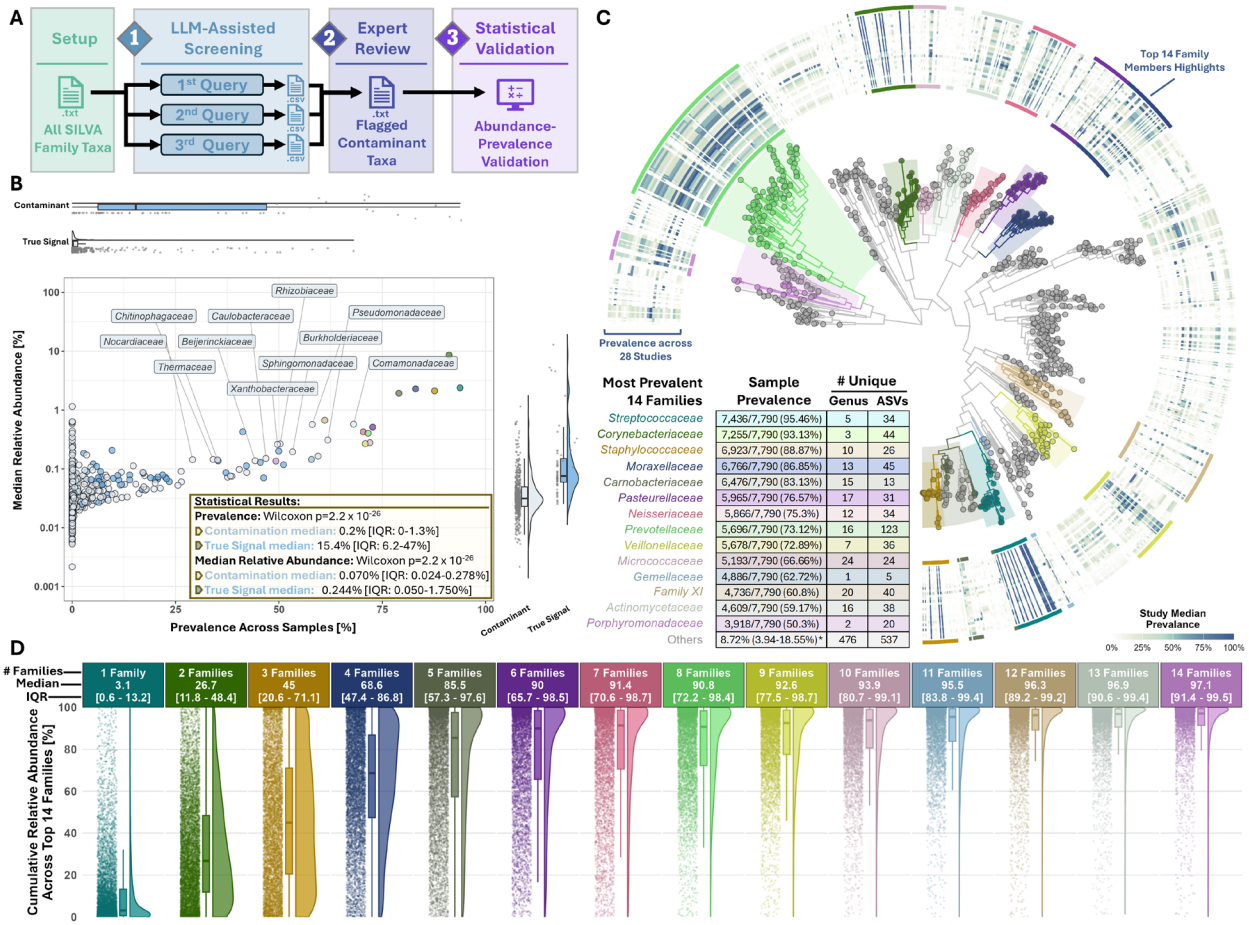


Figure 2. Three-stage contamination assessment and family-level abundance patterns in decontaminated nasopharyngeal microbiome datasets. A. Three-stage contamination identification workflow: (1) LLM-assisted screening of the 596 SILVA-annotated bacterial families detected, (2) expert review to yield initial taxon contamination classifications, and (3) statistical validation via abundance-prevalence analysis, establishing a nasopharyngeal-specific contaminant reference database. **B**. Scatter plot depicting the sample prevalence [%] (x-axis) and median relative abundance [%] (y-axis) of bacterial families colored by contamination status. Axis marginal distributions for contaminant(white) and true signals (i.e., authentic nasopharyngeal families; blue) are shown as boxplots with adjacent jittered data points with Wilcoxon rank-sum statistical analysis results shown in the scatter plot. Selected background families with atypically high prevalence (≥25%) and abundance (≥1%) are labeled. The top 14 prevalent bacterial families (21.9% of the 64 authentic nasopharyngeal families) classified as true signals are separately colored and further described in panels C-D. **C**. Maximum likelihood phylogenetic tree of 1,091 amplicon sequence variants (ASVs) from the representative V3-V4 hypervariable region (**Methods**), with the top 14 prevalent families highlighted by colored rectangular backgrounds organized by phylogenetic clades; overlapping colors

denote shared ancestry. The outer heatmap displays the prevalence of the most-refined taxonomic classification associated with each ASV across the meta-analysis dataset (28 independent studies). **D**. Sample cumulative relative abundance distributions of the top 14 prevalent families, beginning with the most prevalent family (left) and with each subsequent panel showing the accumulation upon adding the next most prevalent family. Each panel displays the family abundance distribution summary statistics (IQR and median) and left-to-right as a jittered data points, boxplot, and density curve.

Characterization of nasopharyngeal NPCSTs

We performed NPCST analysis via hierarchical clustering of 7,790 samples across 28 independent studies, characterizing six distinct and stable NPCSTs (**Figures 3A** and **S5**). These NPCSTs displayed characteristic ecological signatures: NPCST I (*Moraxella*-dominated), NPCST II (*Corynebacterium-Dolosigranulum* co-dominated), NPCST III (*Staphylococcus*-dominated), NPCST IV (*Haemophilus*-dominated), NPCST V (diverse, no dominant genus), and NPCST VI (*Streptococcus*-dominated) (**Figures 3B-C** and **S4**). These NPCST keystone taxa were among the 14 core families (**Figure 2C-D**), and they were highly prevalent (detected in >80% of samples) and abundant within their respective NPCSTs (**Figures 2B**, **3C** and **S4**). NPCSTs explained 52.19% of total variance (PERMANOVA R^2) in community composition, nearly four-fold higher than study-specific batch effects (13.06%) and far exceeding other covariates: HVRs (3.46%), age (3.44%), and sex (0.14%) (**Figure 3B**). This dominant biological signal shows that these six NPCSTs represent fundamental ecological states, each occupying a distinct microbial niche in the nasopharyngeal environment.

These six NPCSTs demonstrated robust reproducibility across multiple validation approaches (see **Supplementary File** for detailed validation) and exhibited community specific taxon abundance patterns (**Figure 3D**). Silhouette score analysis identified six clusters as optimal (**Figures 3E** and **S5**), with leave-one-study-out CV confirming high stability metrics (ARI = 0.711, Jaccard = 0.75, overall accuracy = 0.867). Furthermore, unsupervised learning-based rare biosphere (ulrb) analysis[52] independently validated dominant taxa patterns, revealing 2-17 abundant taxa per NPCST with consistent detection in >95% of samples within each NPCST (**Figures S6-S11**, **Table S4**). Diversity metrics clearly differentiated the NPCSTs: NPCSTs I-IV represented lower-diversity communities dominated by 1-2 genera, whereas NPCSTs V-VI comprised more diverse compositions with higher genus richness and diversity indices (Shannon and Inverse Simpson) (**Figure 3D**).

We developed and validated machine learning models that enable standardized NPCST classification, achieving high median accuracy using both Random Forests (**Figure S12**

and **Table S5**, 96.6%) and Support Vector Machines (**Figure S13** and **Table S5**, 97.2%). These models similarly distinguished NPCSTs by their keystone taxa (**Figures S14** and **S15**) and tended to misclassify the same samples (**Figure S16**, 40-53% across NPCSTs). Since keystone taxa were by far the most important features (**Table S6**), we applied model optimization using Youden's J statistic to reveal clear abundance thresholds for keystone taxa (**Figures S17** and **S18**). We performed Random Forest CV using our external set of 1,573 independently processed samples and achieved consistently high AUCs (>0.91), except for the diverse NPCST V (AUC: 0.853), confirming broad model applicability (**Figure S19**). For samples with ambiguous community membership, our confidence scoring system based on genus abundance and prediction probability improved accuracy to >95% (**Table S7**). These models enable consistent NPCST classification across future nasopharyngeal microbiome 16S studies.

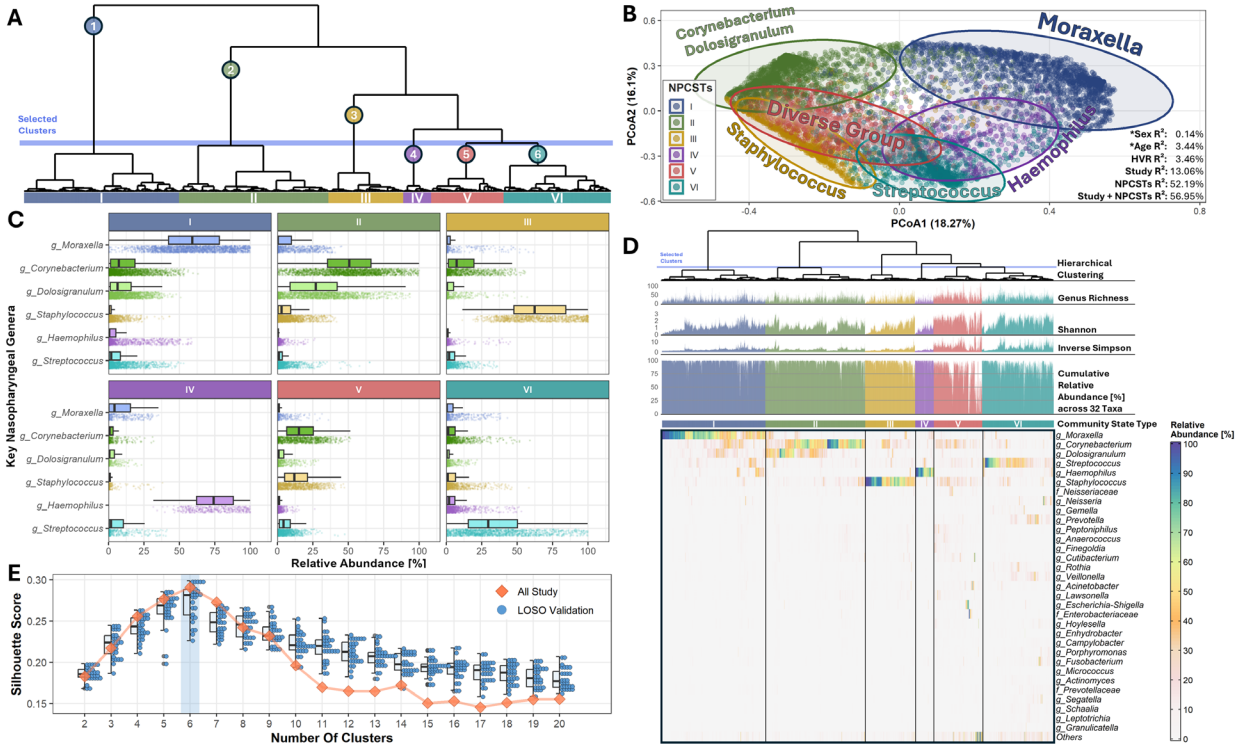


Figure 3. Nasopharyngeal Microbiome Community State Type (NPCST) Analysis. **A.** Hierarchical clustering of nasopharyngeal microbiomes via Ward's Method using pairwise Bray-Curtis dissimilarities. The dendrogram was cut at the blue horizontal line to yield six distinct NPCSTs colored and described further in panels B-E (I=blue, II=green, III=yellow, IV=purple, V=red, VI=teal). **B.** Principal coordinates analysis (PCoA) based on pairwise Bray-Curtis dissimilarities with each NPCST indicated by a colored ellipse representing 95% confidence intervals. The PERMANOVA R^2 of key sample parameters (sex, age, hypervariable region (HVR), study, NPCST) were individually calculated with a combined mode (Study + NPCSTs) displayed. **C.** Relative abundance [%] distributions of key genera

(*g_Moraxella, g_Corynebacterium, g_Dolosigranulum, g_Staphylococcus, g_Streptococcus,* and *g_Haemophilus*) across the six NPCSTs represented by vertically adjacent boxplots and jittered sample data points. **D.** Heatmap displaying relative abundances of 32 urlb-identified taxa with prevalence ≥50% in at least one NPCST (**Figure S6-11**). Above the heatmap top-to-bottom are the sample hierarchical clustering dendrogram, genus richness (i.e., number of genera detected), alpha diversity (Shannon diversity and Inverse Simpson diversity), and cumulative relative abundance of the 32 urlb-identified taxa. For visualization purposes, Inverse Simpson diversity is capped at 10, genus richness at 100, and Shannon diversity at 3. **E.** Silhouette score validation for determining the optimal number of clusters (x-axis), where the leave-one-study-out validation silhouette score distributions are shown as boxplots adjacent to their respective jittered data points (y-axis). The connected orange data points represent silhouette scores calculated using the complete meta-analysis dataset (all 28 independent studies).

Community-State-Specific Patterns in Demographics, Disease Risk, and Microbial Diversity

NPCSTs demonstrated remarkable stability across demographic variables. Analysis of 5,353 cross-sectional samples yielded significant but small effect sizes: age ($\varepsilon^2 = 0.07$, 95% CI: 0.06–0.09) and sex (Cramér's V = 0.07, 95% CI: 0.04–0.09) each explained <7% of community variation (both $p < 0.0001$). The primary demographic pattern emerged in early childhood, with *Moraxella*-dominated (NPCST I) and *Haemophilus*-dominated (NPCST IV) communities preferentially colonizing children under 10 years (**Figure S20A-B**). The otherwise minimal demographic effects underscore the intrinsic stability of NPCSTs as fundamental ecological states of the nasopharyngeal microbiome.

We also observed significant variation in disease risk across NPCSTs when controlling for study effects (**Figure S20** and **Table S8**). Most notably, bacterial infection risk was significantly higher in NPCST VI communities (OR: 1.21; 95% CI: 1.10–1.33; $p < 0.001$) while viral infection risk was significantly lower (OR: 0.90; 95% CI: 0.86–0.94; $p < 0.001$). NPCST IV showed minimal but significant viral susceptibility (OR: 1.05; 95% CI: 1.00–1.10; $p = 0.048$). These findings implicate specific microbiome configurations in respiratory infection susceptibility.

Shannon diversity patterns were intrinsically linked to NPCST identity rather than disease status. Analysis of 5,000-read rarefied samples demonstrated three stable diversity categories: NPCST IV maintained consistently low diversity (<1.5), NPCSTs I–III showed intermediate levels (0–2.5), and NPCSTs V–VI exhibited high diversity (0–3.5) regardless of infection presence (**Figure S20D**). Disease-associated diversity changes within NPCSTs

were minimal and inconsistent, while age contributed negligibly ($R^2 \leq 0.119$; **Figure S20E**). These stable, community-specific patterns have critical implications: diversity metrics must be interpreted within NPCST context.

Co-occurrence Network Analysis

To understand how dominant genera shape microbial interactions within these naturally occurring community states, we constructed NPCST-specific co-occurrence networks to identify whether each NPCST harbors unique patterns of microbial associations beneath the influence of their keystone taxa. Each NPCST contained 25 to 60 taxa with identified associations, resulting in a union of 72 genera from the 626 total genera detected across all samples (**Figure S21** and **Table S9**). SPIEC-EASI analysis revealed different network architectures across NPCSTs: sparse networks in NPCSTs I-III (41-68 edges), minimal connectivity in NPCST IV (1 edge due to *Haemophilus* dominance), and dense networks in the diverse NPCSTs V-VI (79-122 edges) (**Figure 6A-G**). Intriguingly, the keystone taxa characterizing each NPCST formed no reliable associations with other taxa (**Table S9**), likely due to their overwhelming ecological influence and limited reliance on other minor network patterns.

Strikingly, we identified anaerobes as key structural backbones of multiple subdominant community networks despite the aerobic nature of the nasopharynx (**Figure 6H-I**). *Peptoniphilus* and *Anaerococcus* formed an anaerobic hub across nearly all NPCSTs, maintaining 23-81% prevalence and consistent positive associations (**Figure 6H**). These genera, previously understudied due to cultivation challenges[84], demonstrated strong negative associations with *Streptococcus* in NPCSTs I, II, III, and V (**Table S9**). Similarly, the anaerobic pair *Prevotella* and *Veillonella* maintained positive associations across all NPCSTs except IV, with 44-91% prevalence depending on community type.

Aerobic taxa also had a broad presence in the co-occurrence network, forming smaller, NPCST-specific clusters, including *Micrococcus*, *Enhydrobacter*, *Chryseobacterium*, and *Acinetobacter*. These clusters had multiple positive associations and 2-3 fold prevalence variations across NPCSTs, indicating community-specific aerobic niches (**Figure 6H, Figure S21**, **Table S9**). Taken together, these findings show that nasopharyngeal microbiome structures form complex networks of subdominant aerobes and anaerobes that underly the dominant keystone taxa.

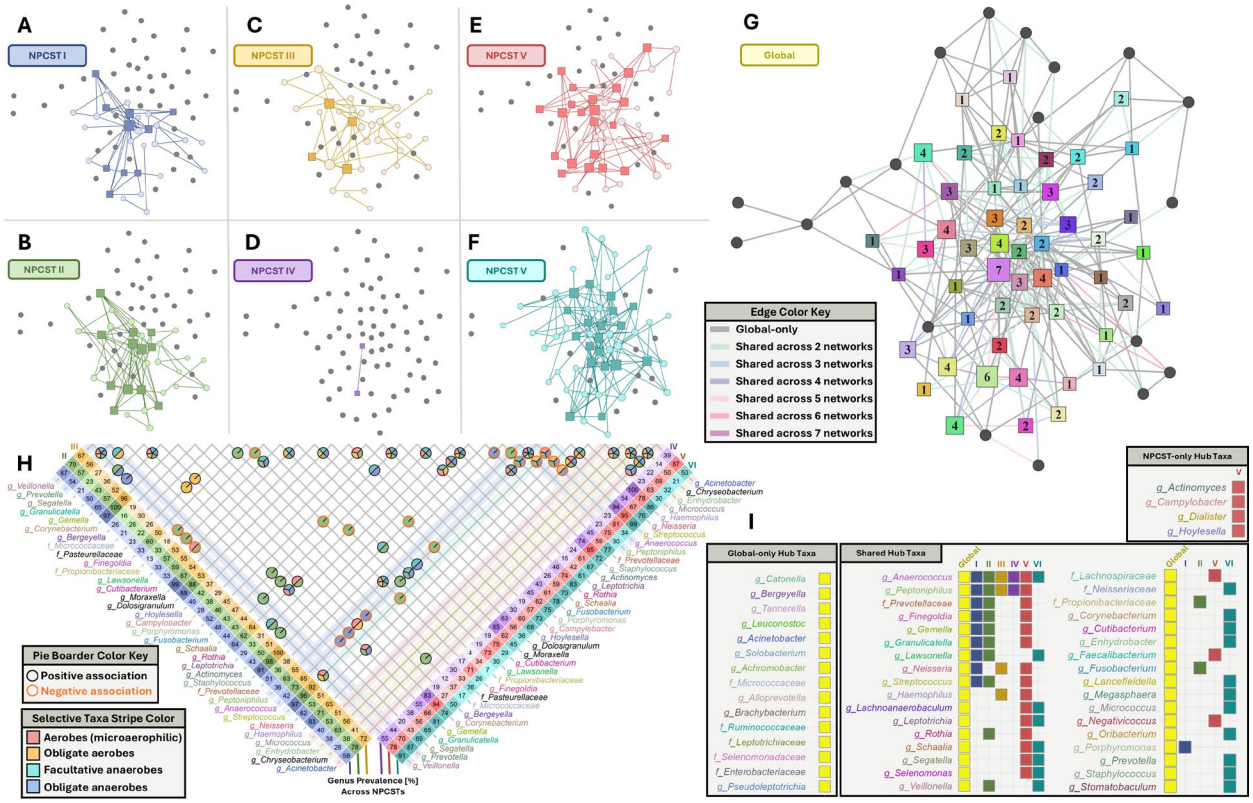


Figure 4. Taxon co-occurrence analysis of nasopharyngeal microbiome community state types (NPCSTs). A-F. NPCST-specific taxon co-occurrence networks derived from SPIEC-EASI analysis. Node colors correspond to NPCST-specific color schemes. Dark-colored squares represent k-core-determined hub taxa within each NPCST (**Methods**), while light-colored circles indicate non-hub taxa with associations. Gray nodes represent taxa present in other networks but lacking associations in the displayed NPCST. As expected, the diverse NPCST IV is mostly represented by gray nodes. **G**. Global co-occurrence network integrating all samples across the six NPCSTs. Squares and circles denote hub taxa and non-hub taxa, respectively, with node size proportional to hub frequency across networks (labeled with number). Nodes (taxa) are also colored by hub category: global-only, NPCST-specific, or shared (global + NPCST). Edge colors indicate network sharing frequency, from gray (single network) to purple (all seven networks). All networks maintain consistent spatial positioning for direct comparison. **H**. Taxon-taxon associations for NPCSTs I–IV, with NPCSTs V–VI shown in Supplementary Figure S21C. Pie charts represent pairwise genera connections, with segmentation indicating associations shared across multiple NPCSTs (color-coded by NPCST). Black and orange circles around the pie charts indicate positive and negative associations, respectively. Notably, all associations maintain consistent directionality across networks. Left and right axes display taxon prevalence percentages across NPCSTs. Taxa are color-coded by oxygen

requirements: aerobes (red), obligate aerobes (orange), facultative anaerobes (teal), and obligate anaerobes (blue). **I.** Legend showing color-coded taxon names and their hub status across network categories.

Prediction of NPCST functions

We next sought to characterize and differentiate the potential NPCST functions. We applied PICRUSt2-SC to predict EC pathway abundances and two approaches for statistical analysis (MaAsLin3 and LinDA), revealing highly consistent functional patterns (**Figure 5A-B** and **Table S10**; 75.4% agreement, CCC = 0.758, Pearson's r = 0.856). Consistent with their distinct community structures (**Figures 3** and **4**), NPCSTs also exhibited clear functional divergence (**Figure 5C**; PERMANOVA $R^2 = 0.37$, $p < 0.001$), with significant pairwise separation (pairwise PERMANOVA R^2 median = 25.94%, IQR: 18.47-28.01%, all FDR-adjusted $p < 0.001$) and consistently large between-NPCST effect sizes (Cohen's d median = 0.90, IQR: 0.70-1.44, all positive values confirming functional coherence) (**Table S11**). NPCST pairs I-IV, II-III, and V-VI tended to cluster along similar functional dimensions along with broad NPCST sample overlap (**Figure 5C**), suggesting NPCSTs could maintain shared metabolic functions despite their distinct taxonomic compositions.

To fully characterize the specific metabolic signatures that distinguish NPCSTs, we applied dynamic percentile filtering and identified 38 key pathways with asymmetric enrichment-depletion patterns across NPCSTs (**Figure 5D** and **Methods**). Low-diversity NPCSTs (I-IV) showed widespread pathway depletion relative to the diverse reference NPCST V (**Figure 5D**), indicating that low-diversity NPCSTs exhibit metabolic streamlining coordinated by their respective dominant keystone taxa. For example, NPCST I (*Moraxella*-dominant) uniquely enriched fatty acid salvage, octane oxidation, and L-tyrosine degradation pathways, while both NPCSTs I and IV (*Haemophilus*-containing) enriched the NAD salvage pathway III, which is known to relate to viral infection immune defense[85]. (**Figure 5D**). In contrast, NPCST II showed depletion across multiple biosynthetic pathways (**Figure 5D**), suggesting that its keystone taxa (*Corynebacterium* and *Dolosigranulum*) form a close symbiotic partnership[86,87] through metabolic complementarity and cross-feeding. These results establish that nasopharyngeal metabolic specializations are distinctly coordinated across NPCSTs.

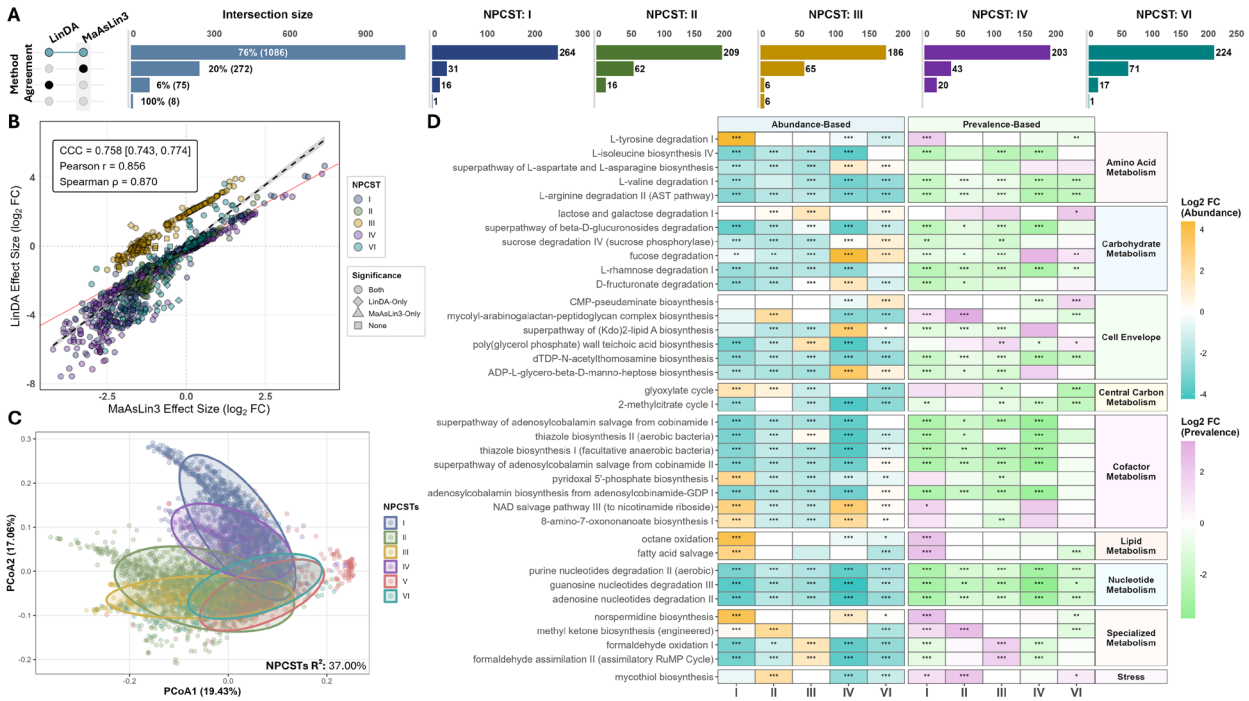


Figure 5. Functional Pathway Analysis of NPCSTs. A. Number of significant Enzyme Commission (EC) functional pathways across NPCSTs with NPCST V (diverse/mixed community) serving as reference. EC functional pathway relative abundances (RAs) were predicted using PICRUSt2-SC followed by two methods to evaluate statistical association with NPCSTs (MaAsLin3 and LinDA). The panel shows left-to-right an upset plot of methodological agreement (i.e., shared/unique statistically significant results), followed by the corresponding total intersection sizes across the NPCST evaluations and stratified plots showing the numbers of significant pathways for each NPCST (I, II, III, IV, VI) compared with NPCST V. **B.** Functional pathway effect size concordance between MaAsLin3 (x-axis) and LinDA (y-axis) for each NPCST comparison with the reference NPCST V, shown as log2 fold-change (FC). Data points (pathways) are colored by NPCST and shaped by significance category, with regression lines and correlation metrics (CCC, Pearson r, Spearman ρ) shown in the top left. **C.** Principal coordinates analysis of pairwise Bray-Curtis dissimilarities calculated using EC functional pathway RAs. Points represent individual samples colored by NPCST with 95% confidence ellipses showing functional differentiation and clustering patterns within and across NPCSTs. **D.** Heatmap depicting significantly enriched and depleted metabolic pathways identified as either shared across multiple NPCSTs or unique to individual NPCSTs, organized by higher-level metabolic categories. Abundance-based coefficients (left; enriched: orange, depleted: blue) and prevalence-based coefficients (right; enriched: purple, depleted: green) are shown with asterisks indicating statistical significance levels (***p<0.001, **p<0.01, * p<0.05).

Nasopharyngeal Microbiome Health Index

NPCSTs harbor distinct keystone taxa, community networks, and functional capacities (**Figures 3-5**), making them essential information for evaluating nasopharyngeal microbiome health. Our observation of differential viral and bacterial susceptibilities across NPCSTs (**Figure S20C-D**) further motivated us to develop a Nasopharyngeal Microbiome Health Index (NMHI). Utilizing the GMWI2 framework[59] as a conceptual foundation, we estimated disease likelihood of the 5,435 cross-sectional nasopharyngeal microbiome samples in our meta-analysis via LASSO-penalized logistic regression with binary presence/absence encoding of taxa abundances (**Methods**). Given that NPCSTs explained substantially more variance than study effects (**Figure 3B**, 53.19% vs. 13.06%), we devised a leave-one-NPCST-out (LONO) strategy for CV in addition to a more traditional 10-fold CV approach. Models were trained across five prevalence thresholds (0-20%) using different microbial feature aggregation techniques (e.g., genus-resolved features only) and classification types (all conditions or viral infections only) (**Methods** and **Table S12**).

We achieved optimal LONO CV results for classification of all conditions using all-taxa and genus-level models with minimal regularization required (**Figure 6A** and **Table S12**; AUCs: 0.886 and 0.847, respectively). We similarly observed strong, stable 10-fold CV performance, with the all-taxa all conditions model performing best at the 0% (AUC=0.900) and 5% (AUC=0.881) prevalence thresholds (**Figure 6A** and **Table S12**). We then selected the 5% prevalence threshold for our NMHI, yielding parsimonious models with ~80% fewer features and only minor reductions in performance (**Figure 6B** and **Table S12**; AUC reduced by ~0.019). The relatively small feature sets that resulted (**Table S12**; all-taxa: 249, genus-level: 88) indicate that these models evaluate nasopharyngeal microbiome health status using core community members rather than rare or species-specific signatures.

Final all-taxa models identified 98 disease-associated and 99 health-promoting taxa, with genus-level models selecting 31 and 37 taxa, respectively (**Figure 6C**). When restricting to high impact taxa (|coefficient| >0.5), we identified 24 disease-associated and 34 health-promoting taxa (**Figure 6C-D**). Some distinct ecological patterns were detected among disease-associated taxa, including *Haemophilus influenzae* with low prevalence of 5-7% and detection among NPCST III, IV, V, and VI; *Corynebacterium kroppenstedtii* with detections across all NPCSTs but 2-3 fold increased prevalence on NPCST II compared to others; and *Achromobacter* with consistent 24-34% prevalence across all NPCSTs (**Figure 6C-D**). These findings revealed one critical insight: First, individual taxa occurrence alone poorly predicts health status, as many taxa appear in both healthy and diseased states and remain rare due to ecological dominance by keystone genera (**Figure S21**). Wilcoxon rank-sum tests confirmed that many taxa identified as health- or disease-associated showed no

significant differential abundance between control and disease groups (**Figures S22-23**). While individual taxa may vary, single-feature approaches miss the community-level reorganization that defines disease states. The NMHI addresses this challenge by binarizing taxa presence/absence and aggregating them into a unified score that captures collective community signals, representing a shift from binary disease classification toward a continuous health spectrum analogous to established clinical metrics such as BMI. -

Next, we applied NMHI threshold optimization to maximize classification accuracy using all samples and within each NPCST (**Figure 6E**). The all-taxa all conditions model achieved excellent balanced accuracy (83%) and AUC (0.91) with an NMHI threshold of 0.2, with NPCST-specific optimization providing minimal improvements in balanced accuracy (79.5-86.4%) (**Figure 6E**). Furthermore, these threshold optimizations yielded large effect sizes (Cohen's d overall: 1.887, NPCST-specific: 1.661-2.012) with statistically significant differences in NMHI (Wilcoxon rank-sum test, all $p<0.001$) (**Figure 6E** and **Table S13**).

To confirm the generalizability of these models, we utilized 699 independent samples from five datasets as an external validation dataset with conditions including SARS-CoV-2 infection, lower tract respiratory infection (LTRI), and critical illness (**Methods**). Using the all-taxa all conditions model with the global NMHI threshold of 0.2 (**Figure 6E**), we observed exceptional discrimination of healthy and 8 diseased populations (**Figure 6F**, median Cohen's d of 2.179). The model also achieved strong balanced accuracy (81.5%) and AUC (0.922), again with our NPCST-specific models minimally improving performance (**Figure 6G**). Together, these results show that our global NMHI evaluation system is highly accurate and robust across NPCSTs and can be generalized to external sample datasets as a respiratory health indicator.

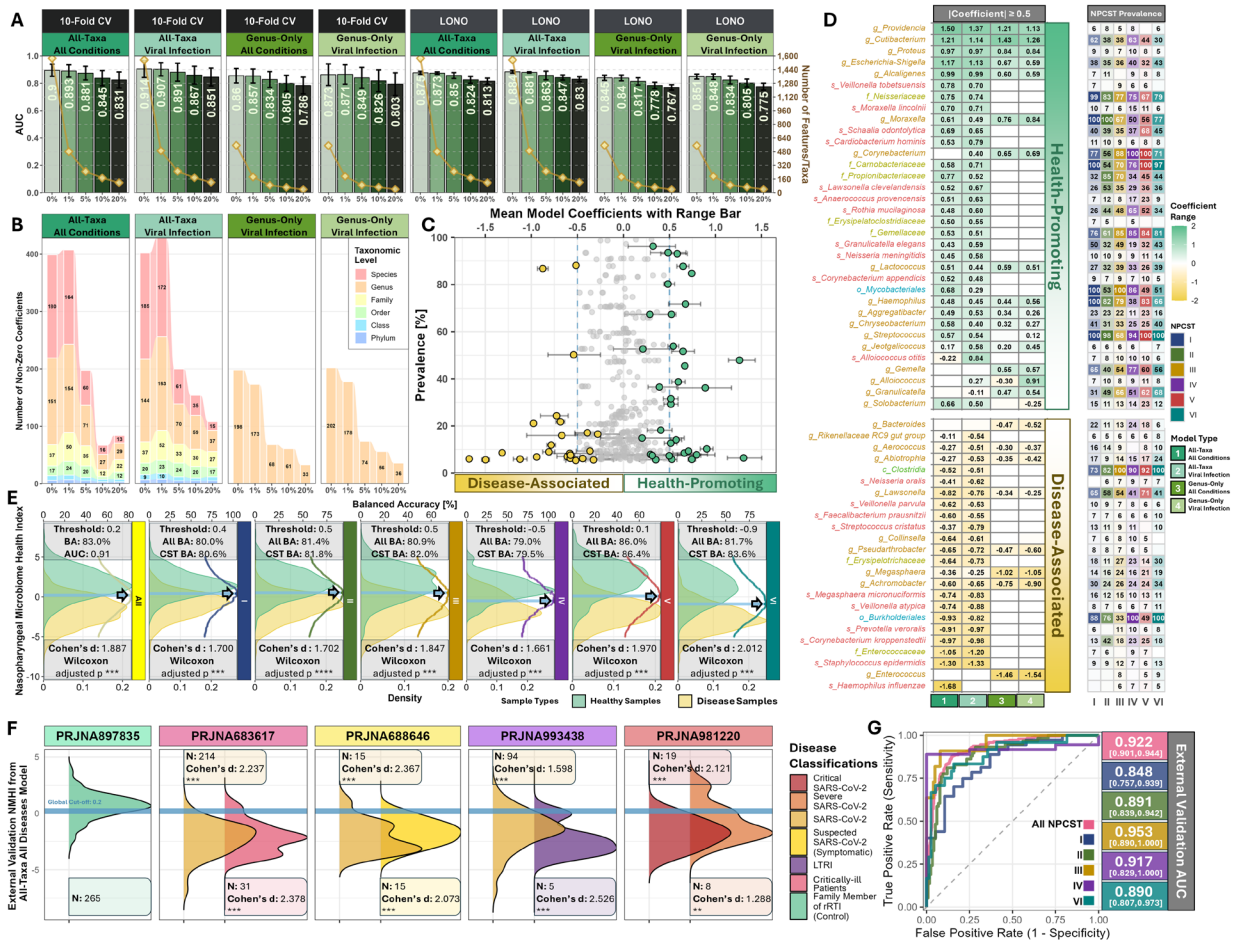


Figure 6. Nasopharyngeal Microbiome Health Index (NMHI) Validation. A. NMHI model performance evaluated using 10-fold cross-validation (CV) and Leave-One-NPCST-Out (LONO) validation across different microbial feature prevalence thresholds (0%, 1%, 5%, 10%, and 20%). Green bars show model median AUC values (left y-axis) with error bars indicating standard deviation across folds (CV) and NPCSTs (LONO). The connected yellow data points indicate the total features used for model training (right y-axis) at each prevalence threshold (x-axis). Each panel shows the results from the combination of validation type (CV or LONO) and the type of model generated, i.e., input microbial feature set used (all taxa or genus-level only) and model classification type (all conditions or viral infections)**. B.** Stacked bar plots showing the numbers of non-zero coefficients in final models trained on the complete dataset, stratified by prevalence threshold and model type. Features are colored by taxonomic resolution (species to phylum) with values ≥10 labeled on the bars. **C.** Mean coefficients (x-axis) and prevalence (y-axis) for each taxon across all four models with error bars indicating the minimum-maximum range. Taxa with an absolute coefficient ≥0.5 (vertical blue dashed lines) in at least one model are highlighted based on whether they are health-promoting (green; positive coefficients) or

disease-associated (yellow; negative coefficients). Gray points represent taxa with absolute coefficients <0.5 across all models. **D.** Heatmaps showing the coefficient magnitudes for high-impact taxa (|coefficient| ≥0.5) across the four models (left) and corresponding prevalence distributions across NPCSTs (right), separately shown for health-promoting (top) and disease-associated (bottom) taxa. **E.** NMHI distributions of healthy controls (green) versus diseased samples (yellow) across all NPCSTs (left) and individual NPCSTs (left-to-right: I-VI). Horizontal dashed lines indicate optimal classification thresholds which are shown above each panel along with measures of balanced accuracy (BA). Balanced accuracy is shown using the global threshold of 0.2 shown in the left-most panel ("All BA") and using the NPCST-specific threshold ("CST BA"). AUC is also shown above the left-most panel. Cohen's d effect sizes and FDR-adjusted Wilcoxon p-values are shown below each panel. **F.** External validation of NMHI across five independent studies, comparing healthy controls with samples from patients with different conditions, including various SARS-CoV-2 infection severities, lower tract respiratory infection (LTRI), and critical illness patients. Cohen's d values and FDR-adjusted Wilcoxon p-values are shown for each comparison with study BioProject accession shown above. **G.** External validation receiver operating characteristic (ROC) curves using global (All) and NPCST-specific thresholds, with AUCs and 95% confidence intervals displayed on the right. An ROC curve and AUC analysis was excluded for NPCST V since it did not contain any health controls. ***p<0.001, **p<0.01.

Discussion

Our study bridges the decades-long gap between nasopharyngeal microbiome research and clinical application through two discoveries: six reproducible community state types explaining 52% of microbial variance and a validated health index (NMHI) that enables quantitative respiratory wellness assessment. These nasopharyngeal community state types, which we term NPCSTs, explain four-fold more variance than technical factors, indicating that they are fundamental biological states beyond demographic and study variables. The NMHI achieves an AUC of 0.92 in external validation, providing clinicians with a continuous wellness metric of the nasopharyngeal microbiome. In addition, we overcame the persistent challenge of contamination in low-biomass samples. Our three-stage decontamination pipeline removed 89% of detected bacterial families as environmental artifacts while preserving biological signals. Together, these contributions establish the analytical standards and practical tools necessary for integrating nasopharyngeal microbiome assessment into respiratory medicine.

Previous nasopharyngeal studies reported inconsistent community clusters (4-13)[20–22] that failed to replicate across cohorts. The six NPCSTs rigorously validated in this study resolve this inconsistency, persisting across age, gender, and study populations. Each NPCST exhibits distinct architecture dominated by keystone genera (*Moraxella*, *Corynebacterium-Dolosigranulum*, *Staphylococcus*, *Haemophilus*, or *Streptococcus*). Remarkably, anaerobic genera *Peptoniphilus* and *Anaerococcus* emerge as network hubs maintaining 23-81% prevalence despite the aerobic environment. The persistence of obligate anaerobes as network hubs suggests they orchestrate microbial interactions that shape community structure, stability, and disease susceptibility. Predictive functional profiling revealed NPCST-specific ecological strategies: *Moraxella*-dominated communities (NPCST I) enrich fatty acid salvage and tyrosine degradation, while *Corynebacterium-Dolosigranulum* partnerships (NPCST II) deplete biosynthetic pathways, suggesting metabolic complementarity. Both NPCSTs showed trends toward reduced bacterial infection risk (NPCST I OR: 0.94, 95% CI: 0.88-1.00; NPCST II OR: 0.93, 95% CI: 0.86-1.00), suggesting protective effects of therapeutic exploration[87]. These findings support developing NPCST-targeted synbiotics, particularly as the respiratory tract offers direct delivery advantages compared to gut and vaginal microbe-centric therapies, such as probiotics, prebiotics, synbiotics and microbiota transplantation[88,89], that must survive gastric acid and digestive enzymes. To enable field-wide adoption, we provide NPCST classification models achieving median AUC of 0.963 across the NPCSTs with the random forest classifier in external validation (n=1,573). These publicly available models enable researchers to consistently assign NPCSTs, establishing the universal framework necessary for comparative studies and clinical translation.

Conflicting diversity patterns in nasopharyngeal studies stemmed from a fundamental oversight: diversity reflects NPCST architecture rather than disease status. SARS-CoV-2 research reported decreased[8,27,74,79], unchanged[9,10,33] and increased[90] Shannon diversity, while studies of LRTI[23], RSV[34] and community-acquired pneumonia[35] similarly showed no consistent alpha diversity patterns. Our analysis reveals that these contradictions arise from three intrinsic diversity tiers that persist regardless of infection status. NPCST IV maintains consistently low diversity (<1.5), NPCSTs I-III show intermediate levels (0-2.5), and NPCSTs V-VI exhibit high diversity (0-3.5), with disease causing minimal changes within each tier. This architectural constraint creates a statistical challenge, where comparing healthy and diseased individuals across NPCSTs with differences in baseline diversity could incorrectly suggest disease association with diversity. Unlike the vaginal microbiome where low diversity[18,19] is a positive health indicator, nasopharyngeal diversity represents community architecture independent of wellness. Therefore, all comparative

analyses require NPCST stratification to avoid spurious conclusions that could misdirect therapeutic development.

Our study has a few limitations. First, 16S rRNA sequencing constrains our analysis to bacterial investigation at genus-level resolution and predicted rather than measured metabolic functions. Metagenomic approaches remain challenging for nasopharyngeal samples with >80% host content[91], but recent validation of an optimal host depletion strategy[91] could help support more metagenomics studies in place of 16S rRNA sequencing. Second, geographic bias toward North America and Asia, combined with overrepresentation of SARS-CoV-2 studies, may limit generalizability to other populations and respiratory conditions. Third, the nasopharyngeal microbiome is known to have seasonal trends[80] (especially during allergy seasons), which will need to be comprehensively studied with more extensive studies. Fourth, the absence of standardized nasopharyngeal sampling protocols across studies introduces variability that our decontamination pipeline cannot fully address. We have identified key methodological inconsistencies that we comprehensively describe in a separate study (Bustos et. al.). Fifth, while our decontamination protocol, which was designed to remove soil and environmental-origin taxa from low-biomass samples, demonstrated robust preservation of biological signals, certain environmental-origin taxa we removed (e.g., *Nocardiaceae*[92]) may be clinically relevant as opportunistic pathogens when studying immunocompromised populations. Depending on the research objectives, the contamination reference database should be refined for such targeted studies. Finally, while we provide publicly available NPCST classification models, clinical implementation requires cross-laboratory standardization. This translation challenge is being actively addressed through ongoing development at Labcorp to implement microbiome-based respiratory assessments. Despite these limitations, our framework provides the foundational tools necessary for advancing nasopharyngeal microbiome research toward routine clinical practice.

Our study establishes a robust foundation for advancing nasopharyngeal microbiome research toward clinical implementation. Our LLM-assisted methodology incorporates human-in-the-loop guidance, aligning with established recommendations for the safe integration of large language models into biomedical research workflows[93]. Beyond nasopharyngeal applications, this decontamination pipeline addresses critical challenges in low-biomass microbiome analysis, for example, extending to tissue microbiota[94] in cancer research. These methodological advances enable mechanistic investigations of biologically significant anaerobic network hubs that emerge only after successful decontamination, providing unprecedented opportunities to study these understudied organisms, particularly their metabolic interactions and pathogen resistance

mechanisms. The distinct functional profiles and disease susceptibilities across NPCSTs support targeted therapeutic development, where single or multi-strain nasopharyngeal probiotics could shift dysbiotic communities toward healthier states, as successfully demonstrated in vaginal microbiome therapeutics[88]. Clinically, the NMHI provides potential utility for monitoring these therapeutic interventions and identifying dysbiosis in PCR-negative symptomatic patients, where our external validation confirmed them as "unhealthy" despite negative PCR tests. Such diagnostic capability fills a critical gap, enabling risk stratification for severe respiratory outcomes and guiding empirical treatment decisions when standard diagnostics provide limited information.

Conclusion

This study developed and validated tools and standards that transform nasopharyngeal sampling into quantitative health assessment. With millions of samples collected annually for respiratory diagnostics, these advances will improve microbial understanding on both population-scale community study and personalized wellness monitoring beyond traditional pathogen detection. The discovery of stable NPCSTs with distinct metabolic strategies and disease susceptibilities opens new therapeutic avenues: NPCST-targeted probiotics, metabolic interventions exploiting salvage pathway dependencies, and personalized respiratory medicine based on community architecture. As respiratory infections remain leading causes of global morbidity, the NMHI provides the missing tool for monitoring intervention efficacy, tracking dysbiosis in vulnerable populations, and identifying at-risk individuals before symptom onset. This work establishes nasopharyngeal microbiome assessment as a clinical reality, positioning it alongside established diagnostics to advance precision respiratory medicine.

Data availability

The data that support the findings of this study are available from multiple sources. Primary data analyzed in this meta-analysis are openly available in the source publications and corresponding NCBI BioProjects (https://www.ncbi.nlm.nih.gov/bioproject/) as listed in **Table 1**. Statistical analysis reports that support the findings of this study are available in the **Supplementary File** of this article. NPCST predictive models are available in Zenodo (doi: 10.5281/zenodo.17068997). Additional supporting data is available on request from the corresponding author.

Author Contributions

Kuncheng Song: Conceptualization; data curation; formal analysis; supervision; project administration; resources; software; validation; visualization; writing - original draft; writing - review & editing. **Hayden N Brochu**: Conceptualization; Data curation; formal analysis; methodology; resources; software; writing - review & editing. **Monica L Bustos**: Data curation; resources; writing - review & editing. **Qimin Zhang**: Data curation; resources; **Crystal R Icenhour**: Funding acquisition; writing - review & editing. **Stanley Letovsky**: Funding acquisition; writing - review & editing. **Lakshmanan K Iyer**: Conceptualization; Supervision; project administration; writing - review & editing.

Funding

This work was supported by the Centers for Disease Control and Prevention (CDC) under awarded contract number 75D30123C17410, BAA 75D301-23-R-72545. This funding supported the development of the Respiratory Epidemiologic Surveillance Network (REaSoN), a surveillance system designed for the early, unbiased detection of novel respiratory pathogens.

Acknowledgements

All authors are current or former employees of Labcorp, a provider of clinical diagnostic services.

We would like to thank the U.S. Centers for Disease Control and Prevention (CDC) for their guidance and expertise throughout the REaSoN project.

References

1. Jung, E. J. *et al*. Comparison of Nasal Swabs, Nasopharyngeal Swabs, and Saliva Samples for the Detection of SARS-CoV-2 and other Respiratory Virus Infections. *Ann Lab Med* **43**, 434–442 (2023).
2. Babady, N. E. The FilmArray® respiratory panel: an automated, broadly multiplexed molecular test for the rapid and accurate detection of respiratory pathogens. *Expert Rev Mol Diagn* **13**, 779–788 (2013).

3. Miller, J. M. *et al.* A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiologya. *Clin Infect Dis* **67**, e1–e94 (2018).
4. Berry, G. J. *et al.* Multicenter Evaluation of the BioFire Respiratory Panel 2.1 (RP2.1) for Detection of SARS-CoV-2 in Nasopharyngeal Swab Samples. *J Clin Microbiol* **60**, e00066-22 (2022).
5. Labcorp. 2019 Novel Coronavirus (COVID-19) With Influenza A, Influenza B and Respiratory Syncytial Virus, NAA. (2025).
6. Tran, P. T., Nduaguba, S. O., Diaby, V., Choi, Y. & Winterstein, A. G. RSV testing practice and positivity by patient demographics in the United States: integrated analyses of MarketScan and NREVSS databases. *Bmc Infect Dis* **22**, 681 (2022).
7. Silk, B. J. *et al.* COVID-19 Surveillance After Expiration of the Public Health Emergency Declaration — United States, May 11, 2023. *MMWR Morb. Mortal. Wkly. Rep.* **72**, 523–528 (2023).
8. Martin-Castaño, B. *et al.* The relationship between gut and nasopharyngeal microbiome composition can predict the severity of COVID-19. *Elife* **13**, RP95292 (2025).
9. Dietl, B. *et al.* Differential nasopharyngeal microbiota patterns: A Comparative Study of Pneumococcal Pneumonia, COVID-19, and Healthy Adults. *J Infect* 106589 (2025) doi:10.1016/j.jinf.2025.106589.
10. Claus, J. *et al.* Nasopharyngeal microbiome composition by SARS-CoV-2 presence and severity. *Sci Rep-uk* **15**, (2025).

11. Braun, T. *et al*. SARS-CoV-2 does not have a strong effect on the nasopharyngeal microbial composition. *Sci Rep-uk* **11**, 8922 (2021).

12. O'Sullivan, D. M. *et al*. An inter-laboratory study to investigate the impact of the bioinformatics component on microbiome analysis using mock communities. *Sci Rep-uk* **11**, 10590 (2021).

13. Salter, S. J. *et al*. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *Bmc Biol* **12**, 87 (2014).

14. Rajar, P. *et al*. Microbial DNA extraction of high-host content and low biomass samples: Optimized protocol for nasopharynx metagenomic studies. *Front Microbiol* **13**, 1038120 (2022).

15. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, (2018).

16. Griffard-Smith, R. *et al*. micRoclean: an R package for decontaminating low-biomass 16S-rRNA microbiome data. *Front Bioinform* **5**, (2025).

17. Yoon, S. M., Ki, C.-S. & Song, J. S. CleanSeqU algorithm for decontamination of catheterized urine 16S rRNA sequencing data. *Sci Rep-uk* **15**, (2025).

18. France, M. T. *et al*. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome* **8**, 166 (2020).

19. Ravel, J. *et al*. Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4680–4687 (2011).

20. Castro-Severyn, J. *et al.* Impact of seasonal variation on the oral and nasopharyngeal microbiome in school-aged children: the school MicroBE initiative. *Msystems* e00467-25 (2025) doi:10.1128/msystems.00467-25.

21. Hernandez-Leyva, A. J. *et al.* Upper and lower airway microbiota across infancy and childhood. *Pediatr Res* (2025) doi:10.1038/s41390-025-03942-0.

22. Lappan, R. *et al.* A microbiome case-control study of recurrent acute otitis media identified potentially protective bacterial genera. *Bmc Microbiol* **18**, 13 (2018).

23. Claassen-Weitz, S. *et al.* Nasopharyngeal Microbiota in South African Infants With Lower Respiratory Tract Infection: A Nested Case-Control Study of the Drakenstein Child Health Study. *Clin Infect Dis* ciaf184 (2025) doi:10.1093/cid/ciaf184.

24. Crovetto, F. *et al.* Nasopharyngeal microbiota profiling of pregnant women with SARS-CoV-2 infection. *Sci Rep-uk* **12**, 13404 (2022).

25. Kelly, M. S. *et al.* Non-diphtheriae Corynebacterium species are associated with decreased risk of pneumococcal colonization during infancy. *ISME J* **16**, 655–665 (2022).

26. Merenstein, C. *et al.* Signatures of COVID-19 Severity and Immune Response in the Respiratory Tract Microbiome. *Mbio* **12**, e01777-21 (2021).

27. Kim, J. G. *et al.* The salivary and nasopharyngeal microbiomes are associated with SARS-CoV-2 infection and disease severity. *J Med Virol* **95**, e28445 (2023).

28. Tchoupou Saha, O. L. F. *et al.* Profile of the Nasopharyngeal Microbiota Affecting the Clinical Course in COVID-19 Patients. *Front Microbiol* **13**, 871627 (2022).

29. Paulo, A. C., Lança, J., Almeida, S. T., Hilty, M. & Sá-Leão, R. The upper respiratory tract microbiota of healthy adults is affected by Streptococcus pneumoniae carriage, smoking habits, and contact with children. *Microbiome* **11**, 199 (2023).

30. De Steenhuijsen Piters, W. A. A. *et al.* Early-life viral infections are associated with disadvantageous immune and microbiota profiles and recurrent respiratory infections. *Nat Microbiol* **7**, 224–237 (2022).

31. Hurst, J. H. *et al.* Age-Related Changes in the Nasopharyngeal Microbiome Are Associated With Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection and Symptoms Among Children, Adolescents, and Young Adults. *Clin Infect Dis* **75**, e928–e937 (2022).

32. Aydin, M. *et al.* The Rhinobiome of Exacerbated Wheezers and Asthmatics: Insights From a German Pediatric Exacerbation Network. *Front Allergy* **2**, 667562 (2021).

33. Žuštra, A. *et al.* Longitudinal dynamics of the nasopharyngeal microbiome in response to SARS-CoV-2 Omicron variant and HIV infection in Kenyan women and their children. *Msystems* **10**, e01568-24 (2025).

34. Kristensen, M. *et al.* The respiratory microbiome is linked to the severity of RSV infections and the persistence of symptoms in children. *Cell Reports Medicine* **5**, 101836 (2024).

35. Haak, B. W. *et al.* Bacterial and Viral Respiratory Tract Microbiota and Host Characteristics in Adults With Lower Respiratory Tract Infections: A Case-Control Study. *Clin Infect Dis* **74**, 776–784 (2022).

36. Ventero, M. P. *et al.* Nasopharyngeal Microbiota as an early severity biomarker in COVID-19 hospitalised patients. *J Infection* **84**, 329–336 (2022).

37. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* n71 (2021) doi:10.1136/bmj.n71.

38. National Center for Biotechnology Information. PubMed. (1996).

39. Elsevier. Embase. (1974).

40. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).

41. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).

42. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013).

43. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* **27**, 105–117 (2019).

44. Anthropic. Claude. (2022).

45. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

46. Sievers, F. & Higgins, D. G. Clustal Omega. *CP in Bioinformatics* **48**, (2014).

47. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2021).

48. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

49. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *CP in Bioinformatics* **69**, e96 (2020).

50. Oksanen, J. *et al. Vegan: Community Ecology Package*. (2024).

51. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*. (2023).

52. Pascoal, F., Branco, P., Torgo, L., Costa, R. & Magalhães, C. Definition of the microbial rare biosphere through unsupervised machine learning. *Commun Biol* **8**, 544 (2025).

53. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput Biol* **11**, e1004226 (2015).

54. Csárdi, G. *et al. Igraph: Network Analysis and Visualization in R*. (2025). doi:10.5281/zenodo.7682609.

55. Wright, R. J. & Langille, M. G. I. PICRUSt2-SC: an update to the reference database used for functional prediction within PICRUSt2. *Bioinformatics* **41**, btaf269 (2025).

56. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res* **48**, D445–D453 (2020).

57. Nickols, W. A. *et al.* MaAsLin 3: Refining and extending generalized multivariable linear models for meta-omic association discovery. Preprint at https://doi.org/10.1101/2024.12.13.628459 (2024).

58. Zhou, H., Chen, J., Zhang, X. & He, K. *LinDA: Linear Model for Differential Abundance Analysis*. (2025).

59. Chang, D. *et al.* Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles. *Nat Commun* **15**, 7447 (2024).

60. Kassambara, A. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. (2023).

61. Signorell, A. *DescTools: Tools for Descriptive Statistics*. (2024).

62. Elff, M. *Mclogit: Multinomial Logit Models, with or without Random Effects or Overdispersion*. (2022).

63. Lenth, R. V. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. (2024).

64. Edouard, S. *et al.* The nasopharyngeal microbiota in patients with viral respiratory tract infections is enriched in bacterial pathogens. *Eur J Clin Microbiol* **37**, 1725–1733 (2018).

65. De Boeck, I. *et al.* Anterior Nares Diversity and Pathobionts Represent Sinus Microbiome in Chronic Rhinosinusitis. *mSphere* **4**, e00532-19 (2019).

66. Jörissen, J. *et al.* Case-Control Microbiome Study of Chronic Otitis Media with Effusion in Children Points at Streptococcus salivarius as a Pathobiont-Inhibiting Species. *Msystems* **6**, 10.1128/msystems.00056-21 (2021).

67. Pittet, L. F. *et al. Chlamydia pneumoniae* and *Mycoplasma pneumoniae* in children with cystic fibrosis: impact on bacterial respiratory microbiota diversity. *Pathog Dis* **79**, ftaa074 (2021).

68. Shetty, S. A. *et al.* Associations and recovery dynamics of the nasopharyngeal microbiota during influenza-like illness in the aging population. *Sci Rep-uk* **12**, 1915 (2022).

69. Thorsen, J. *et al.* The airway microbiota of neonates colonized with asthma-associated pathogenic bacteria. *Nat Commun* **14**, 6668 (2023).

70. Bozan, G. *et al.* Analysis of Intestinal and Nasopharyngeal Microbiota of Children with Meningococcemia in Pediatric Intensive Care Unit: INMACS-PICU Study. *Diagnostics* **13**, 1984 (2023).

71. Ruiz-Tagle, C. *et al.* Reduced microbial diversity of the nasopharyngeal microbiome in household contacts with latent tuberculosis infection. *Sci Rep-uk* **13**, 7301 (2023).

72. Htun, K. S. *et al.* Microbiome dataset from the upper respiratory tract of patients living with HIV, HIV/TB and TB from Myanmar. *Data Brief* **21**, 354–357 (2018).

73. Bowers, J. R. *et al.* Genomic Analyses of Acute Flaccid Myelitis Cases among a Cluster in Arizona Provide Further Evidence of Enterovirus D68 Role. *Mbio* **10**, e02262-18 (2019).

74. Ventero, M. P. *et al.* Nasopharyngeal Microbial Communities of Patients Infected With SARS-CoV-2 That Developed COVID-19. *Front Microbiol* **12**, 637430 (2021).

75. Galeeva, J. S. *et al.* Microbial Signatures in COVID-19: Distinguishing Mild and Severe Disease via Gut Microbiota. *Biomedicines* **12**, 996 (2024).

76. Karampatsas, K. *et al.* Gastrointestinal, vaginal, nasopharyngeal, and breast milk microbiota profiles and breast milk metabolomic changes in Gambian infants over the first two months of lactation: A prospective cohort study. *Medicine* **101**, e31419 (2022).

77. Ferrari, L. *et al.* Nasopharyngeal Bacterial Microbiota Composition and SARS-CoV-2 IgG Antibody Maintenance in Asymptomatic/Paucisymptomatic Subjects. *Front Cell Infect Mi* **12**, 882302 (2022).

78. Koenen, M. H. *et al.* Mycoplasma pneumoniae carriage in children with recurrent respiratory tract infections is associated with a less diverse and altered microbiota. *eBioMedicine* **98**, 104868 (2023).

79. Romani, L. *et al.* Exploring nasopharyngeal microbiota profile in children affected by SARS-CoV-2 infection. *Microbiol Spectr* **12**, e03009-23 (2024).

80. Odendaal, M.-L. *et al.* Host and environmental factors shape upper airway microbiota and respiratory health across the human lifespan. *Cell* S0092867424007682 (2024) doi:10.1016/j.cell.2024.07.008.

81. Sun, D., Tang, L., Long, K., Sun, W. & Zhang, Z. Bacterial contamination in the different parts of household air conditioners: a comprehensive evaluation from Chengdu, Southwest China. *Front Public Health* **12**, 1429626 (2024).

82. Pati, A. *et al.* Complete genome sequence of Oceanithermus profundus type strain (506T). *Stand. Genomic Sci.* **4**, 210–220 (2011).

83. Kachuei, R., Emami, M., Mirnejad, R. & Khoobdel, M. Diversity and frequency of Nocardia spp. in the soil of Isfahan province, Iran. *Asian Pac J Trop Bio* **2**, 474–478 (2012).

84. Yu, K. *et al.* Interactions between bacteria in the human nasopharynx: a scoping review. *The Lancet Microbe* 101062 (2025) doi:10.1016/j.lanmic.2024.101062.

85. Shang, J., Smith, M. R., Anmangandla, A. & Lin, H. NAD+-consuming enzymes in immune defense against viral infection. *Biochem J* **478**, 4071–4092 (2021).

86. Cisneros, M. *et al.* Synergistic inhibition of pneumococcal growth by *Dolosigranulum pigrum* and *Corynebacterium pseudodiphtheriticum:* insights into nasopharyngeal microbial interactions. *Microbiol Spectr* **13**, (2025).

87. Stubbendieck, R. M., Hurst, J. H. & Kelly, M. S. Dolosigranulum pigrum: A promising nasal probiotic candidate. *PLoS Pathog* **20**, e1011955 (2024).

88. Ravel, J. *et al.* Impact of a multi-strain L. crispatus-based vaginal synbiotic on the vaginal microbiome: a randomized placebo-controlled trial. *NPJ Biofilms Microbiomes* **11**, 158 (2025).

89. Daisley, B. A. & Allen-Vercoe, E. Microbes as medicine. *Annals of the New York Academy of Sciences* **1541**, 63–82 (2024).

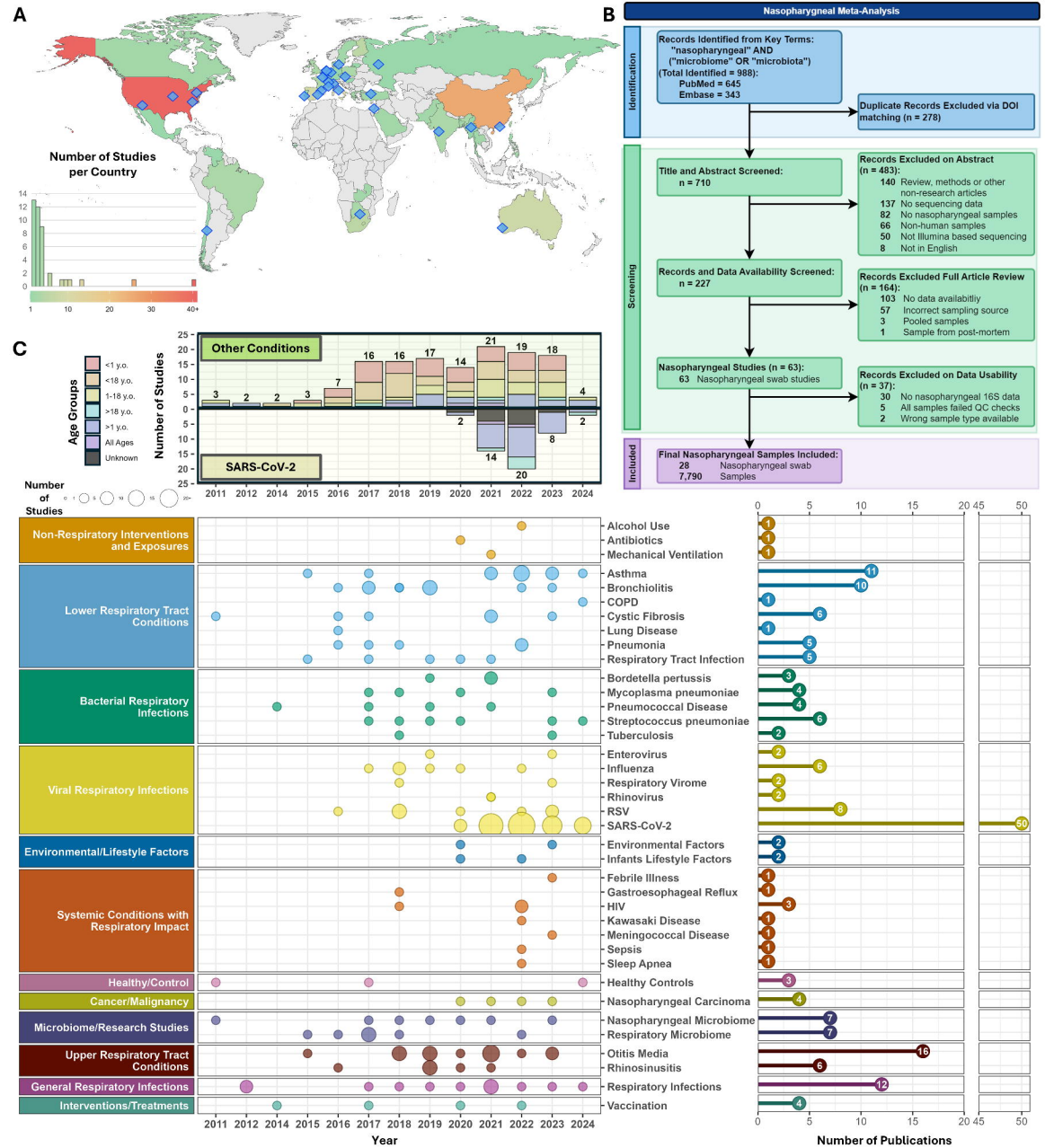
90. Galeana-Cadena, D. *et al.* Microbiome in the nasopharynx: Insights into the impact of COVID-19 severity. *Heliyon* **10**, e31562 (2024).

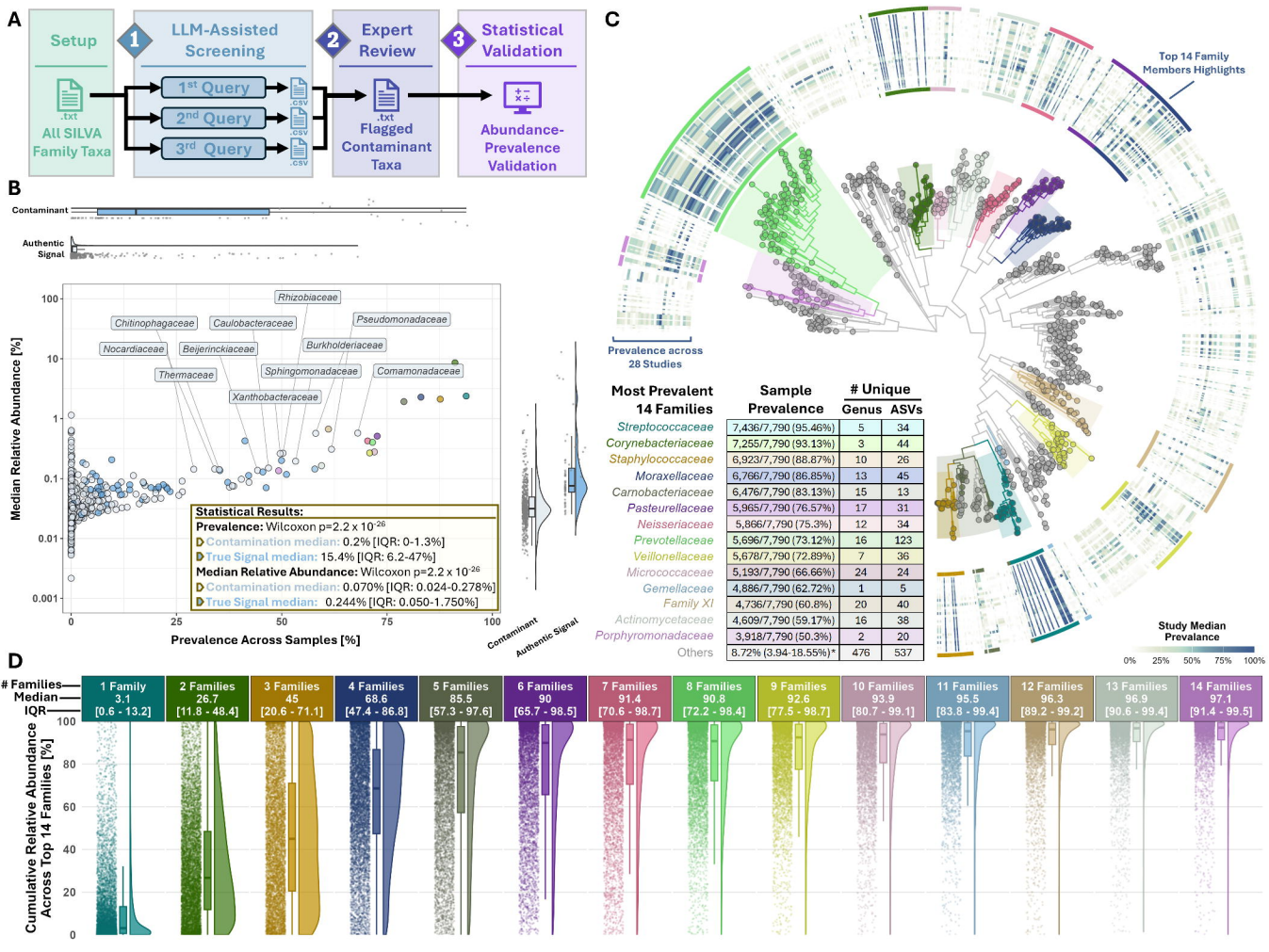
91. Wang, C. *et al.* Benefits and challenges of host depletion methods in profiling the upper and lower respiratory microbiome. *Npj Biofilms Microbi* **11**, (2025).

92. Wilson, J. W. Nocardiosis: updates and clinical overview. *Mayo Clin Proc* **87**, 403–407 (2012).

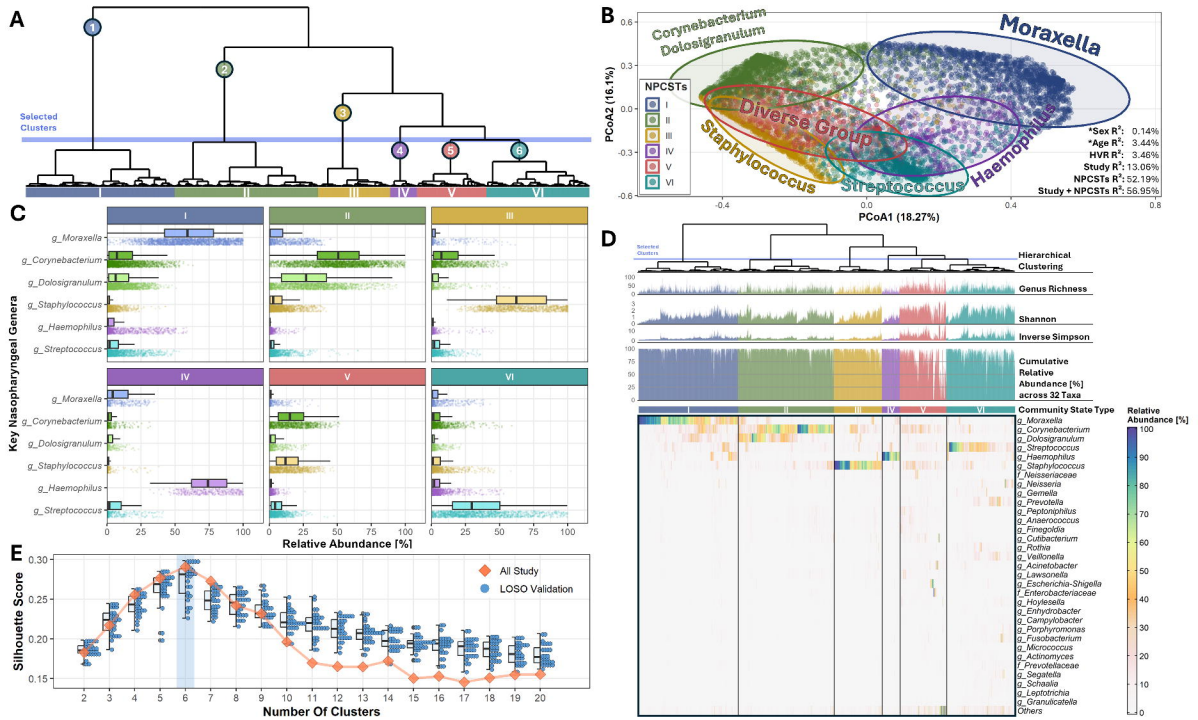
93. Lu, Z. *et al.* Large language models in biomedicine and health: current research landscape and future directions. *J Am Med Inform Assoc* **31**, 1801–1811 (2024).

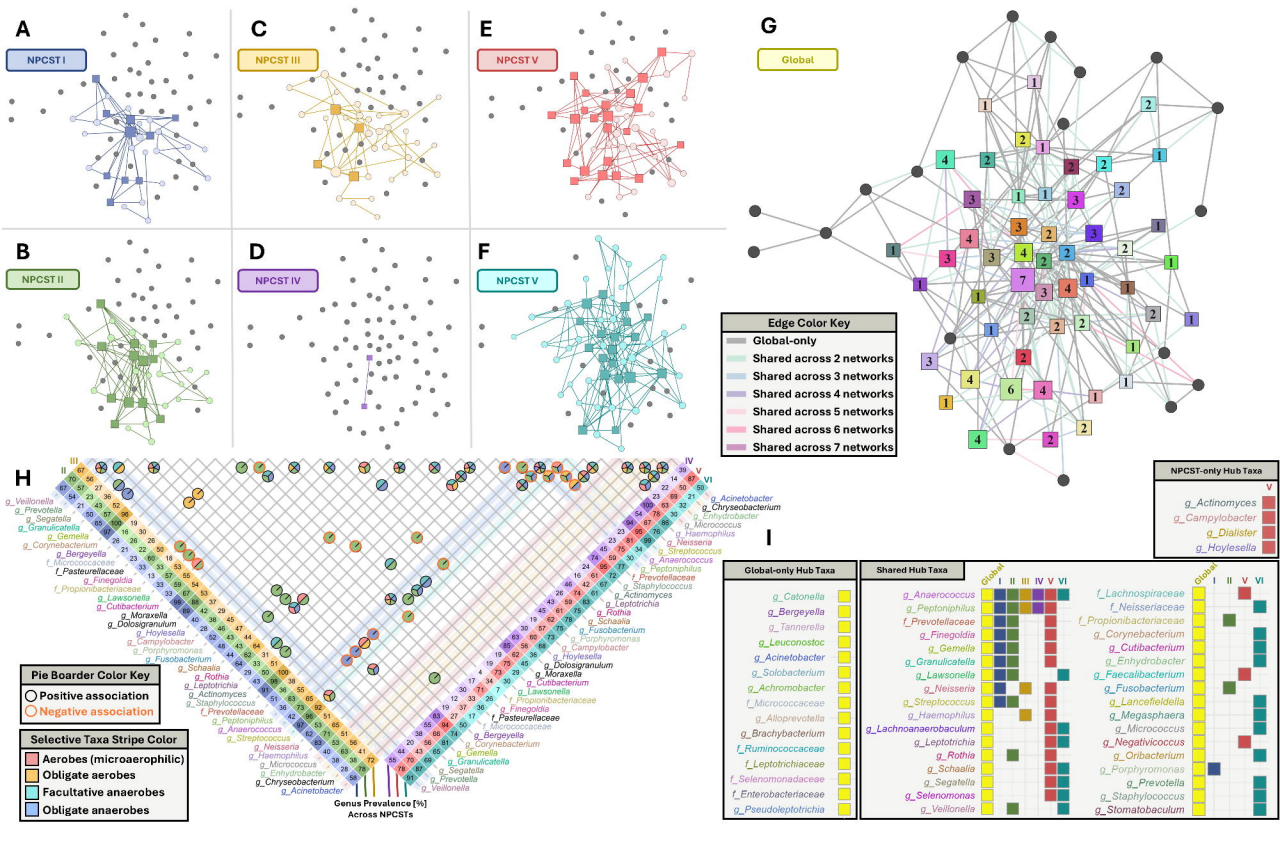
94. Sirasani, J. P., Gardner, C., Jung, G., Lee, H. & Ahn, T.-H. Bioinformatic approaches to blood and tissue microbiome analyses: challenges and perspectives. *Brief Bioinform* **26**, bbaf176 (2025).

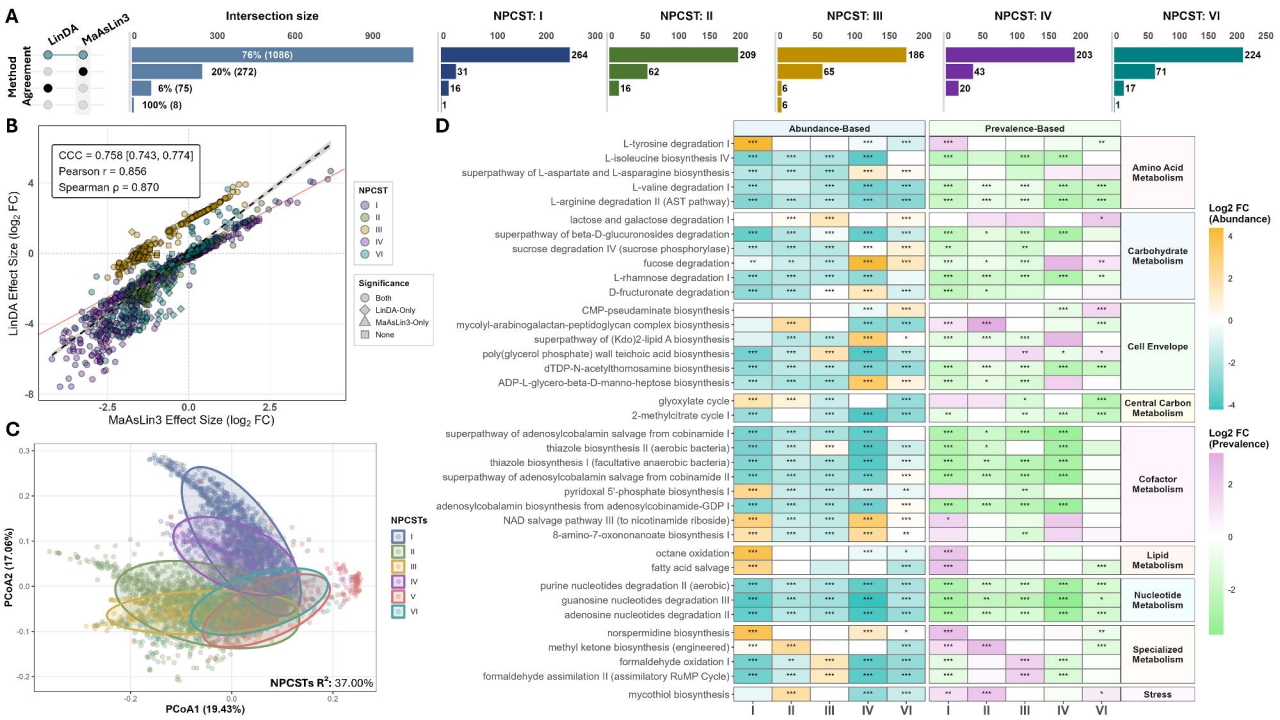

A
Number of Studies per Country
B
Nasopharyngeal Meta-Analysis
Identification
Records Identified from Key Terms: "nasopharyngeal" AND ("microbiome" OR "microbiota") (Total Identified = 988): PubMed = 645 Embase = 343
Duplicate Records Excluded via DOI matching (n = 278)
Screening
Title and Abstract Screened: n = 710
Records Excluded on Abstract (n = 483):
140 Review, methods or other non-research articles
137 No sequencing data
82 No nasopharyngeal samples
66 Non-human samples
50 Not Illumina based sequencing
8 Not in English
Records and Data Availability Screened: n = 227
Records Excluded Full Article Review (n = 164):
103 No data availabitliy
57 Incorrect sampling source
3 Pooled samples
1 Sample from post-mortem
Nasopharyngeal Studies (n = 63):
63 Nasopharyngeal swab studies
Records Excluded on Data Usability (n = 37):
30 No nasopharyngeal 16S data
5 All samples failed QC checks
2 Wrong sample type available
Included
Final Nasopharyngeal Samples Included:
28 Nasopharyngeal swab
7,790 Samples
C
Age Groups
<1 y.o.
<18 y.o.
1-18 y.o.
>18 y.o.
>1 y.o.
All Ages
Unknown
Number of Studies
Other Conditions
SARS-CoV-2
Number of Studies
Non-Respiratory Interventions and Exposures
Alcohol Use
Antibiotics
Mechanical Ventilation
Lower Respiratory Tract Conditions
Asthma
Bronchiolitis
COPD
Cystic Fibrosis
Lung Disease
Pneumonia
Respiratory Tract Infection
Bacterial Respiratory Infections
Bordetella pertussis
Mycoplasma pneumoniae
Pneumococcal Disease
Streptococcus pneumoniae
Tuberculosis
Viral Respiratory Infections
Enterovirus
Influenza
Respiratory Virome
Rhinovirus
RSV
SARS-CoV-2
Environmental/Lifestyle Factors
Environmental Factors
Infants Lifestyle Factors
Systemic Conditions with Respiratory Impact
Febrile Illness
Gastroesophageal Reflux
HIV
Kawasaki Disease
Meningococcal Disease
Sepsis
Sleep Apnea
Healthy/Control
Healthy Controls
Cancer/Malignancy
Nasopharyngeal Carcinoma
Microbiome/Research Studies
Nasopharyngeal Microbiome
Respiratory Microbiome
Upper Respiratory Tract Conditions
Otitis Media
Rhinosinusitis
General Respiratory Infections
Respiratory Infections
Interventions/Treatments
Vaccination
Year
Number of Publications

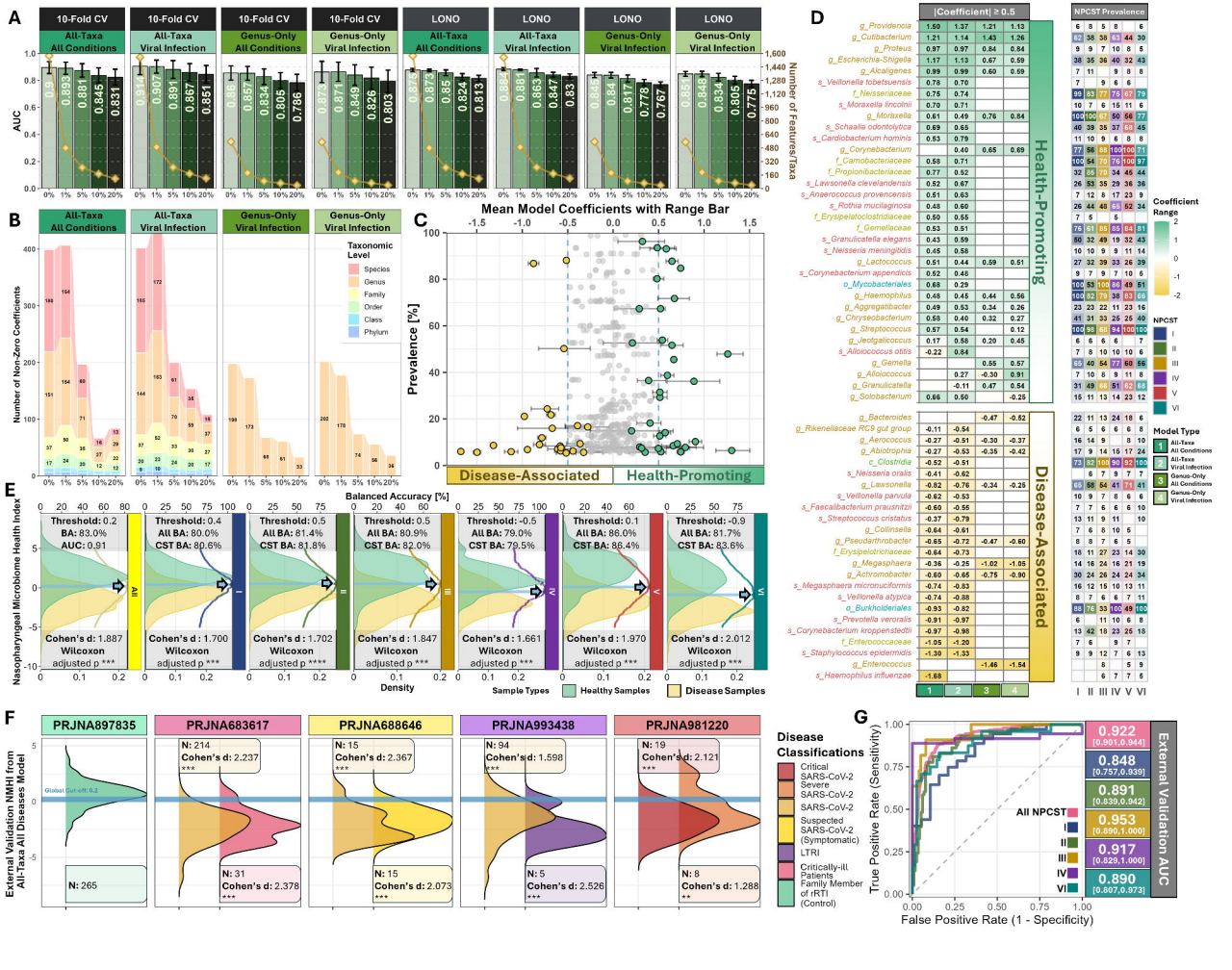


Most Prevalent 14 Families	Sample Prevalence	# Unique Genus	# Unique ASVs
Streptococcaceae	7,436/7,790 (95.46%)	5	34
Corynebacteriaceae	7,255/7,790 (93.13%)	3	44
Staphylococcaceae	6,923/7,790 (88.87%)	10	26
Moraxellaceae	6,766/7,790 (86.85%)	13	45
Carnobacteriaceae	6,476/7,790 (83.13%)	15	13
Pasteurellaceae	5,965/7,790 (76.57%)	17	31
Neisseriaceae	5,866/7,790 (75.3%)	12	34
Prevotellaceae	5,696/7,790 (73.12%)	16	123
Veillonellaceae	5,678/7,790 (72.89%)	7	36
Micrococcaceae	5,193/7,790 (66.66%)	24	24
Gemellaceae	4,886/7,790 (62.72%)	1	5
Family XI	4,736/7,790 (60.8%)	20	40
Actinomycetaceae	4,609/7,790 (59.17%)	16	38
Porphyromonadaceae	3,918/7,790 (50.3%)	2	20
Others	8.72% (3.94-18.55%)*	476	537


A
Selected Clusters
I
II
III
IV
V
VI
B
Corynebacterium
Dolosigranuium
Moraxella
Diverse Group
Staphylococcus
Streptococcus
Haemophilus
NPCSTs
PCoA2 (16.1%)
PCoA1 (18.27%)
*Sex R²: 0.14%
*Age R²: 3.44%
HVR R²: 3.46%
Study R²: 13.06%
NPCSTs R²: 52.19%
Study + NPCSTs R²: 56.95%
C
Key Nasopharyngeal Genera
g_Moraxella
g_Corynebacterium
g_Dolosigranulum
g_Staphylococcus
g_Haemophilus
g_Streptococcus
Relative Abundance [%]
D
Hierarchical Clustering
Genus Richness
Shannon
Inverse Simpson
Cumulative Relative Abundance [%] across 32 Taxa
Community State Type
Relative Abundance [%]
Others
E
Silhouette Score
Number Of Clusters
All Study
LOSO Validation


A
NPCST I
B
NPCST II
C
NPCST III
D
NPCST IV
E
NPCST V
F
NPCST V
G
Global
Edge Color Key
Global-only
Shared across 2 networks
Shared across 3 networks
Shared across 4 networks
Shared across 5 networks
Shared across 6 networks
Shared across 7 networks
H
Pie Boarder Color Key
Positive association
Negative association
Selective Taxa Stripe Color
Aerobes (microaerophilic)
Obligate aerobes
Facultative anaerobes
Obligate anaerobes
Genus Prevalence [%]
Across NPCSTs
I
NPCST-only Hub Taxa
g_Actinomyces
g_Campylobacter
g_Dialister
g_Hoylesella
Global-only Hub Taxa
g_Catonella
g_Bergeyella
g_Tannerella
g_Leuconostoc
g_Acinetobacter
g_Solobacterium
g_Achromobacter
f_Micrococcaceae
g_Alloprevotella
g_Brachybacterium
f_Ruminococcaceae
f_Leptotrichiaceae
f_Selenomonadaceae
f_Enterobacteriaceae
g_Pseudoleptotrichia
Shared Hub Taxa
Global
I
II
III
IV
V
VI
g_Anaerococcus
g_Peptoniphilus
f_Prevotellaceae
g_Finegoldia
g_Gemella
g_Granulicatella
g_Lawsonella
g_Neisseria
g_Streptococcus
g_Haemophilus
g_Leptotrichia
g_Rothia
g_Schaalia
g_Fusobacterium
g_Porphyromonas
g_Campylobacter
g_Hoylesella
g_Moraxella
g_Dolosigranulum
g_Cutibacterium
f_Propionibacteriaceae
f_Pasteurellaceae
f_Lachnospiraceae
f_Neisseriaceae
f_Propionibacteriaceae
g_Corynebacterium
g_Cutibacterium
g_Enhydrobacter
g_Faecalibacterium
g_Fusobacterium
g_Lancefieldella
g_Megasphaera
g_Micrococcus
g_Negativicoccus
g_Oribacterium
g_Porphyromonas
g_Prevotella
g_Staphylococcus
g_Stomatobaculum

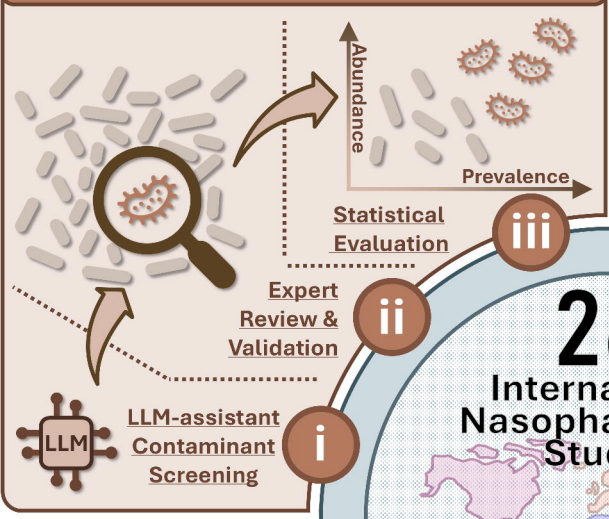

A
LinDA
MaAsLin3
Method Agreement
Intersection size
76% (1086)
20% (272)
6% (75)
100% (8)
NPCST: I
264
31
16
1
NPCST: II
209
62
16
NPCST: III
186
65
6
6
NPCST: IV
203
43
20
NPCST: VI
224
71
17
1
B
CCC = 0.758 [0.743, 0.774]
Pearson r = 0.856
Spearman ρ = 0.870
LinDA Effect Size (log2 FC)
MaAsLin3 Effect Size (log2 FC)
NPCST
I
II
III
IV
VI
Significance
Both
LinDA-Only
MaAsLin3-Only
None
C
PCoA2 (17.06%)
PCoA1 (19.43%)
NPCSTs
NPCSTs R2: 37.00%
D
Abundance-Based
Prevalence-Based
L-tyrosine degradation I
L-isoleucine biosynthesis IV
superpathway of L-aspartate and L-asparagine biosynthesis
L-valine degradation I
L-arginine degradation II (AST pathway)
lactose and galactose degradation I
superpathway of beta-D-glucuronosides degradation
sucrose degradation IV (sucrose phosphorylase)
fucose degradation
L-rhamnose degradation I
D-fructuronate degradation
CMP-pseudaminate biosynthesis
mycolyl-arabinogalactan-peptidoglycan complex biosynthesis
superpathway of (Kdo)2-lipid A biosynthesis
poly(glycerol phosphate) wall teichoic acid biosynthesis
dTDP-N-acetylthomosamine biosynthesis
ADP-L-glycero-beta-D-manno-heptose biosynthesis
glyoxylate cycle
2-methylcitrate cycle I
superpathway of adenosylcobalamin salvage from cobinamide I
thiazole biosynthesis II (aerobic bacteria)
thiazole biosynthesis I (facultative anaerobic bacteria)
superpathway of adenosylcobalamin salvage from cobinamide II
pyridoxal 5'-phosphate biosynthesis I
adenosylcobalamin biosynthesis from adenosylcobinamide-GDP I
NAD salvage pathway III (to nicotinamide riboside)
8-amino-7-oxononanoate biosynthesis I
octane oxidation
fatty acid salvage
purine nucleotides degradation II (aerobic)
guanosine nucleotides degradation III
adenosine nucleotides degradation II
norspermidine biosynthesis
methyl ketone biosynthesis (engineered)
formaldehyde oxidation I
formaldehyde assimilation II (assimilatory RuMP Cycle)
mycothiol biosynthesis
Amino Acid Metabolism
Carbohydrate Metabolism
Cell Envelope
Central Carbon Metabolism
Cofactor Metabolism
Lipid Metabolism
Nucleotide Metabolism
Specialized Metabolism
Stress
Log2 FC (Abundance)
Log2 FC (Prevalence)
I
II
III
IV
VI



A
10-Fold CV
All-Taxa All Conditions
All-Taxa Viral Infection
Genus-Only All Conditions
Genus-Only Viral Infection
LONO
AUC
Number of Features/Taxa
B
Number of Non-Zero Coefficients
Taxonomic Level
Species
Genus
Family
Order
Class
Phylum
C
Mean Model Coefficients with Range Bar
Prevalence [%]
Disease-Associated
Health-Promoting
D
|Coefficient| ≥ 0.5
NPCST Prevalence
Health-Promoting
Disease-Associated
Coefficient Range
NPCST
Model Type
E
Balanced Accuracy [%]
Nasopharyngeal Microbiome Health Index
Threshold: 0.2
BA: 83.0%
AUC: 0.91
Cohen's d: 1.887
Threshold: 0.4
All BA: 80.0%
CST BA: 80.6%
Cohen's d : 1.700
Threshold: 0.5
All BA: 81.4%
CST BA: 81.8%
Cohen's d : 1.702
Threshold: 0.5
All BA: 80.9%
CST BA: 82.0%
Cohen's d : 1.847
Threshold: -0.5
All BA: 79.0%
CST BA: 79.5%
Cohen's d : 1.661
Threshold: 0.1
All BA: 86.0%
CST BA: 86.4%
Cohen's d : 1.970
Threshold: -0.9
All BA: 81.7%
CST BA: 83.6%
Cohen's d : 2.012
Wilcoxon
Density
Sample Types
Healthy Samples
Disease Samples
F
PRJNA897835
PRJNA683617
PRJNA688646
PRJNA993438
PRJNA981220
External Validation NMHI from All-Taxa All Diseases Model
Global Cut-off: 0.2
N: 265
N: 214
Cohen's d: 2.237
N: 31
Cohen's d: 2.378
N: 15
Cohen's d: 2.367
N: 15
Cohen's d: 2.073
N: 94
Cohen's d: 1.598
N: 5
Cohen's d: 2.526
N: 19
Cohen's d: 2.121
N: 8
Cohen's d: 1.288
Disease Classifications
Critical SARS-CoV-2
Severe SARS-CoV-2
SARS-CoV-2
Suspected SARS-CoV-2 (Symptomatic)
LTRI
Critically-ill Patients
Family Member of rRTI (Control)
G
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
All NPCST
External Validation AUC
0.922 [0.901,0.944]
0.848 [0.757,0.939]
0.891 [0.839,0.942]
0.953 [0.890,1.000]
0.917 [0.829,1.000]
0.890 [0.807,0.973]

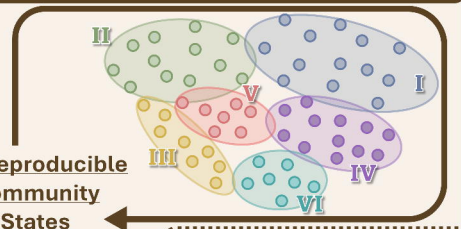
Three Stage Decontamination Protocol:

LLM screening, expert curation, and statistical validation establishes nasopharyngeal-specific contamination reference

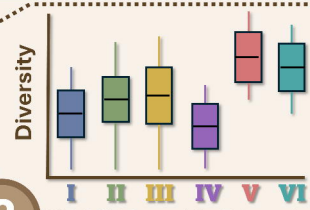


Nasopharyngeal Community State Types (NPCSTs):

Reproducible classification with distinct diversity, co-occurrence patterns and metabolic signatures

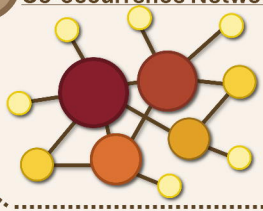


1 Six Reproducible Community States



2 NPCST-Specific Diversity

3 NPCST-Specific Co-occurrence Networks



4 NPCST-Specific Functional Differentiation

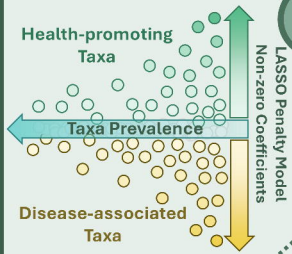
Differential Abundance Compared to NPCST V

	I	II	III	IV	VI
Function 1				↓	↑
Function 2	↓	↑↑		↓↓	↓↓
Function 3	↑↑↑			↓	↓
Function 4	↑↑↑			↑	↓
Function 5	↓↓↓	↓	↑	↓↓↓	↓↓

28 International Nasopharyngeal Studies

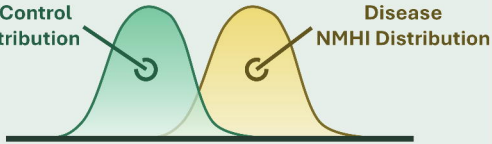
Uniformly Re-processed High-Quality Samples 7,790

Identification of Key Taxa



I

II NMHI Validation



NMHI Review and Validation

Nasopharyngeal Microbiome Health Index (NMHI):

Clinical translation framework shifts from binary disease detection to continuous health monitoring